What war has done to science in Russia p. 632

Why property wildfire risks are rising pp. 638 & 702

AI that anticipates emergent illicit drugs wins NOMIS & *Science* Award p. 656

Science

$15
10 NOVEMBER 2023
science.org
AAAS

CHIP LASER

Powerful mode-locked lasers miniaturized to chip scale p. 708

EDITORIAL

Global fight against HIV is at risk

Wafaa M. El-Sadr is the director of ICAP at Columbia University and professor of Epidemiology and Medicine at the Mailman School of Public Health, Columbia University, New York, NY, USA. wme1@columbia.edu

Myron S. Cohen is the director of the Institute for Global Health and Infectious Diseases and professor of Medicine, Microbiology and Immunology and Epidemiology at the University of North Carolina, Chapel Hill, NC, USA. mscohen@med.unc.edu

Will the United States remain committed to a landmark health program it started 20 years ago that saved more than 25 million lives around the world? For a younger generation of physicians, nurses, and researchers today, the depth of despair wrought by the HIV/AIDS epidemic at that time is almost unimaginable. Even more profoundly, for countries in Africa, the epidemic presented an existential threat. Without access to antiretroviral therapy or efficacious prevention tools, new infections continued unfettered, and people with HIV/AIDS faced near-certain death. Yet, somehow out of this anguish came a ray of hope. In 2003, President George W. Bush announced the launch of the President's Emergency Plan for AIDS Relief (PEPFAR), committing billions of dollars to fight the epidemic in the poorest countries around the world. The initiative has been a shining example of global collaboration in the face of adversity. Yet, as of this September, the US Congress has yet to reauthorize PEPFAR. The Biden administration, individuals from both sides of the political divide, and former President Bush himself have urged Congress to support PEPFAR's lifesaving work. The scientific and public health communities must do the same before it is too late.

> "...PEPFAR has transformed health systems around the world..."

PEPFAR, by all accounts, has been a resounding success. At its inception, only 50,000 persons living with HIV in sub-Saharan Africa had access to treatment. Today, this number has reached over 20 million. This achievement altered the face of the continent, saving families and reviving communities. Beyond these direct effects, PEPFAR has transformed health systems around the world by working with numerous governments, in-country partners and civil society groups, and global organizations including the Global Fund to Fight AIDS, Tuberculosis and Malaria. These partnerships trained hundreds of thousands of health workers, equipped thousands of laboratories and pharmacies, and established numerous surveillance and data systems. Indeed, countries used these resources to jump-start their COVID-19 responses and save millions more lives.

PEPFAR's work is not finished. There were 1.3 million new HIV infections and 630,000 HIV-related deaths reported in 2022. Population surveys in African countries consistently show that children, young people, and men lag in accessing HIV prevention and treatment services. In addition, stigmatized, at-risk populations have been left behind, including men who have sex with men, transgender persons, people who inject drugs, and sex workers. In response, PEPFAR has charted a way forward that is anchored in fully engaging affected communities and expanding access to high-quality prevention and treatment programs that are tailored to their needs.

Beyond PEPFAR's undisputed benefits in the more than 50 countries that it supports, the program's accomplishments offer lessons to guide the path to a successful HIV response in the United States. Rapid adoption of innovations will improve the quality and reach of services. These include the scale-up of self-testing, differentiated service-delivery methods to meet the needs and preferences of specific populations, and the dynamic use of up-to-date data to direct resources and inspire action in a timely manner. And expanding the scope of work for community health workers and peer educators, which includes people living with HIV and persons from affected communities, will help reach those facing stigma and who are in denial.

From day one, PEPFAR has been distinguished by the bipartisan support it has garnered over successive administrations, which has led to its enthusiastic reauthorization by Congress every 5 years for over two decades. Congressional support has rested on the program's remarkable achievements as well as its standing as exemplary of the US global health commitment. Yet, current reauthorization is at an impasse because of misperceptions and inaccurate assertions that have no bearing whatsoever on PEPFAR's purpose and work. Some conservative voices contend that PEPFAR funds support access to abortions, assertions that PEPFAR staff and public health leaders repeatedly affirm are groundless. Other concerns point to PEPFAR language regarding groups that scientific data have shown to be at-risk for HIV and whose members need HIV prevention and treatment services, including transgender people and sex workers. These concerns were not raised in previous reauthorizations.

Scientists, health care experts, global health leaders, and advocates must stand up at this critical moment, learn more about PEPFAR, and articulate with clarity what the program does and does not do. Now is the time to meet and share information with lawmakers and Congressional staff. Write letters. Speak out. The lives of millions hang in the balance.

–Wafaa M. El-Sadr* and Myron S. Cohen

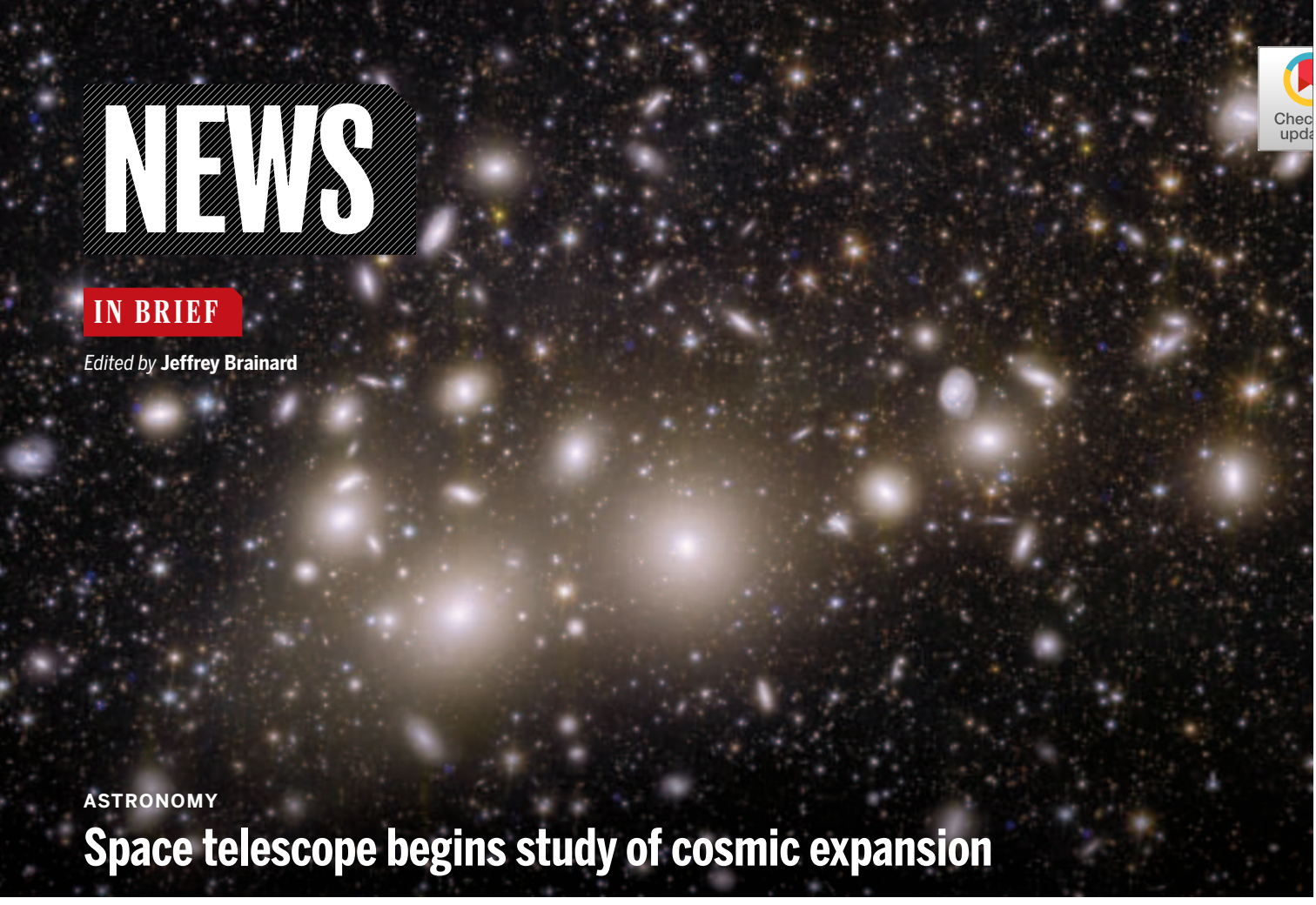
*ICAP has received funding from PEPFAR through the Centers for Disease Control and Prevention and the United States Agency for International Development.

Published online 2 November 2023; 10.1126/science.adm6975

NEWS

IN BRIEF

Edited by **Jeffrey Brainard**



ASTRONOMY

Space telescope begins study of cosmic expansion

The European Space Agency this week released the first pictures of galaxies taken by its orbiting telescope, Euclid, which could help researchers understand the dark components that make up 95% of the universe. The new telescope's image of the Perseus Cluster (above), one of the most massive structures in the universe, shows 1000 of its galaxies 240 million light-years from Earth. Behind them are 100,000 more distant ones, some as far away as 10 billion light-years. Over its 6-year mission, Euclid, launched in July, is expected to take 30,000 such images, cataloging 1 billion galaxies across one-third of the sky. Researchers will use them to create the biggest ever 3D map of the universe, spanning three-quarters of its history. By recording the galaxies' locations, clustering, and shapes, scientists hope for insights about two of the thorniest problems in astrophysics: the nature of dark matter, which is needed to keep spinning galaxies from flying apart, and dark energy, which accelerates the universe's expansion.

AI safety research expands

POLICY | The United States and United Kingdom said last week they will create research institutes to help develop artificial intelligence (AI) that is safe and trustworthy. The announcements coincided with a summit at which 28 countries agreed to cooperate on studying and mitigating risks that AI will worsen cyberattacks, bioweapons, and disinformation. The new U.S. AI Safety Institute, led by the National Institute of Standards and Technology (NIST), will consult academic and industry groups on ways to measure AI capabilities. The consultations are also meant to support NIST's work to help watermark AI-generated content, reduce harmful algorithmic discrimination, and preserve privacy. The U.K. version, based in London, will test the safety of AI models, complementing existing international efforts. Some skeptics of last week's summit criticized a lack of binding commitments.

Pseudoscience in India's schools

CURRICULUM | Educational authorities in India have drawn criticism for publishing school teaching materials containing pseudoscientific claims drawn from Hindu mythology. The educational modules—focused on Chandrayaan-3, the Indian space mission that in September landed a probe on the Moon—mention *vimānas*, which are flying palaces or chariots described in Sanskrit epics. The modules' creator, the National Council of Educational Research and Training, withdrew them last month after criticism but then republished them after the Ministry of Education defended them, saying, "Mythology and philosophy put forward ideas, and ideas lead to innovation and research." Critics also pointed to factual errors in the teaching content and said it reflected cultural nationalism. Members

of Prime Minister Narendra Modi's government have previously been criticized for asserting that ancient Hindu stories describe accomplishments in surgery, stem cell therapies, and relativity.

Birds to lose people's names

BIOLOGY | The American Ornithological Society last week said it will change the common names of at least 80 U.S. and Canadian bird species whose monikers are "clouded by racism and misogyny." For example, the Scott's oriole (*Icterus parisorum*), found in the southwestern United States, is named for Winfield Scott, a Civil War general who oversaw the forced expulsion of Indigenous people from their lands. The society, which maintains an official list of bird names, says additional renamings may follow. It has no plans to change species' Latin scientific names, some of which also refer to people. This "is probably one of the most controversial subjects we've had to deal with in the bird world for a very long time," says Sushma Reddy, an ornithologist at the University of Minnesota, Twin Cities, and the society's secretary. To develop new species names, the society plans to consult people from diverse academic backgrounds.

Funders push open-access growth

PUBLISHING | The Coalition S group of funders, which 5 years ago required its grantees to begin making their research free to read when published, last week unveiled a new vision for largely voluntary steps to further shake up scientific publishing. The Towards Responsible Publishing manifesto supports, for example, a shift toward funding open-access publication through "diamond" approaches, in which research institutions and funders, not authors or subscribers, pay for publishing costs. The draft plan suggests funders support these and other recommendations by giving preference to grantees who choose them. Coalition S plans to gather comments and finalize the proposal by June 2024. In 2022, its funders, most of which are based in Europe, financed the research in less than 4% of published papers. But some credit it for helping raise the portion of papers that are open access to 56% of those published in 2022.

Science camp gets $200 million

PHILANTHROPY | A 64-year-old summer research program for high-achieving high school students has received a $200 million windfall from the estate of an alumnus, one of the largest gifts ever for K-12 education. "We were shocked and thrilled," says Frank Steslow, CEO of the Summer Science Program (SSP), which this year enrolled 204 rising high school seniors in an intensive, 6-week lab course in astrophysics, biochemistry, or genomics at one of five U.S. universities. The unexpected bequest came from Franklin Antonio, a 1969 SSP graduate who later co-founded chipmaking giant Qualcomm and died last year at the age of 69. Steslow says SSP, which had less than $7 million in total assets in 2021, hopes to use the unrestricted gift, announced last month, to enroll a larger and more diverse pool of students. It also plans to offer more scholarships (tuition is now $8400) and perhaps build and run its own network of remotely operated telescopes.

Rigorous psych studies replicate

OPEN SCIENCE | Ninety percent of a sample of rigorously conducted behavioral science studies could be replicated, a study has found, pointing to a possible solution to the discipline's replication crisis. The rigorous practices include studying large samples; publicly posting study instruments, such as questionnaires; and announcing hypotheses and methods in advance. (Authors of such preregistered studies commit to publishing all results, however dull or dazzling.) Researchers had previously found that only 30% to 70% of behavioral studies in which an experimental intervention had a positive result could be replicated, and the interventions yield much smaller effects in replication attempts than in the original work. Some have argued that low rates of replication could stem from the inherent variability of human behavior. But how the research is done can make a difference, researchers reported this week in *Nature Human Behaviour*: Four groups submitted 13 positive results from such studies to be replicated four times each—and 47 of these 52 replication attempts confirmed the positive result and found effect sizes similar to those in the original studies. The analysis is the first that followed experiments from conception through to independent replication, the authors say.

A stoplight parrotfish feeds by scraping algae off coral.

PHOTO: CLAUDIO CONTRERAS/NPL/MINDEN

TAXONOMY

Revised map of fish lineage illuminates family ties

Researchers are flocking to download a revised tree of life for ray-finned fishes, which account for half the living vertebrate species and 97% of all living fish. In the first comprehensive synthesis of all the classification work of these animals, ichthyologists Thomas Near of Yale University and Christine Thacker of the Los Angeles Natural History Museum covered most of the fish species consumed by humans and used as pets. Unlike earlier phylogenies that were based primarily on morphology, theirs incorporates extensive DNA data. Among the monograph's fresh insights: Instead of being a distinct family as other trees indicated, the many species of colorful, reef-dwelling parrotfish are members of the wrasse fish family and developed jaws modified for scraping algae off reefs. The 800-page monograph has been downloaded almost 10,000 times since its release in September; by comparison, the American Society of Ichthyologists and Herpetologists only lists 2400 members—and some of them study only amphibians and reptiles.

In January 2024, the U.S. government plans to sell off its helium reserve in Texas.

NATURAL RESOURCES

Impending U.S. helium sale worries researchers

Fearing supply disruptions, industry calls for delay to unloading of federal reserve

By **Julia Rosen**

After serving as the world's largest provider of helium for nearly a century, the U.S. government is close to bowing out of the market. Aiming to recoup costs, it began emptying its vast underground reservoir of the gas in the 1990s and has now put the whole federal helium system up for sale. Early next year, the government plans to unveil bids from potential buyers of the Amarillo, Texas, facility, which includes a distribution pipeline and the porous rock formation containing the remaining stores of helium—a substance used in everything from MRI machines to quantum computers to party balloons.

But few are celebrating.

Last month, the Compressed Gas Association (CGA), which represents industrial gas companies, together with groups representing the semiconductor, medical, and aerospace sectors, warned the sale could disrupt supplies of the indispensable element. Scientists worry it will send already stiff prices soaring, especially after the government ended a program that sold helium to federally funded researchers at a discount. And a bipartisan group of senators wrote to the Biden administration in late October to call for a pause to the sale. Last week, in an email to potential bidders, the government delayed the sale from mid-November to 25 January 2024, but for reasons unrelated to supply concerns. "It's sort of like it's on this administrative train to be sold and we don't know how to stop it," says CGA President Rich Gottwald.

In addition to being lighter than air, helium is staunchly unreactive, making it useful for manufacturing sensitive products such as computer chips. It can also be chilled to –269°C, making it essential for cooling quantum computers as well as the superconducting magnets inside MRI machines.

But on Earth helium is scarce. It forms through the radioactive decay of uranium and thorium deep in the crust and eventually seeps out and escapes into space. Some, however, gets trapped in the same rocks as natural gas deposits. As a result, helium extraction is typically a byproduct of fossil fuel production, which has made resource development spotty.

The federal government began to hoard the gas 100 years ago to ensure supplies for military dirigibles and later, the space program. (A ban on helium exports in the 1930s forced Germany to fill the Hindenburg with explosive hydrogen instead of inert helium.) By the 1990s, Congress decided it had spent too much on the stockpile and directed the Bureau of Land Management (BLM) to start selling it off. The agency did so at below market rates, which some say discouraged private companies from expanding their own helium operations.

Today, helium is produced primarily in just four countries—the United States, Qatar, Algeria, and Russia—and is distributed by a handful of multinational companies. The limited number of suppliers together with rising demand has resulted in a roller coaster market. In 2022, for instance, an explosion at a Russian plant and a 5-month shutdown at the U.S. reserve caused a shortage and drove high-grade helium prices to roughly $11 per cubic meter, according to the U.S. Geological Survey—up from about $6 a decade earlier.

As the only large-scale storage facility in the world, the federal helium reserve can release the gas to buffer market shocks, just as the U.S. Strategic Oil Reserve helped

combat rising gasoline prices last year. Selling it "potentially adds more uncertainty and volatility to the market," says Janie Chermak, an economist at the University of New Mexico. Chermak co-authored a 2010 report by the National Academies of Sciences, Engineering, and Medicine that concluded selling the reserve "is not in the best interest of U.S. taxpayers or the country."

Scientists often find themselves on the sharp end of helium price spikes and, as small consumers, sometimes can't get any helium at all. "You have to be worrying about the supply at all times," says Sophia Hayes, a chemist at Washington University in St. Louis who uses nuclear magnetic resonance machines to study the structure of materials. Liquid helium is needed to cool the machines' magnets and, in some cases, the samples themselves.

Hayes fears the federal sale will only make things worse for scientists. Mark Elsesser, director of public affairs for the American Physical Society, helped craft a 2021 bill that would have extended the discount program for researchers for 8 years and bought time for labs to ramp up helium recycling capabilities. The bill went nowhere, but the senators' letter urges the White House to revisit the idea as part of the sale.

Gottwald says companies fear supply problems when the reserve changes hands. Production could be disrupted by a maintenance backlog and the need for the new owner to secure rights of way for 680 kilometers of pipeline, which will not automatically transfer. There is also the matter of negotiating a new contract with the Cliffside Helium Enrichment Unit—a privately owned facility that is necessary for processing helium at the reserve. (BLM and the General Services Administration, which is handling the sale, did not respond to requests for comment.)

A hiccup at the reserve could affect about 10% of annual global production, says Maura Garvey, president of Intelligas Consulting. "That will cause us to go into tight supply all over again." CGA is asking for a 3-year delay, enough time to get 45 million cubic meters of already purchased crude helium out of the ground, according to Garvey. The government plans to sell the remaining 51 million cubic meters alongside the reserve in January 2024.

In a few years, Garvey expects the helium shortage to ease and prices to drop. Russia and Qatar have large plants coming online, and a major new project in Wyoming recently won investor backing. High prices have also encouraged a growing number of operations that target helium as the primary product—not as a byproduct of natural gas.

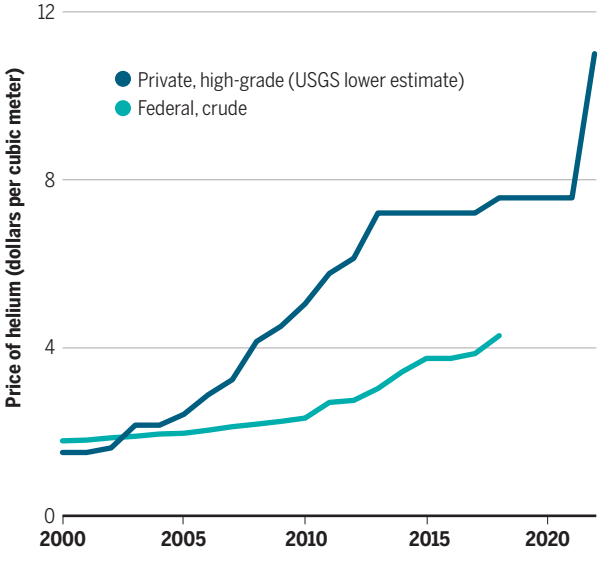
Canadian company Avanti Helium is among the helium prospectors. CEO Chris Bakker says the geopolitical risks of the helium market prompted his team to look for more North American sources. Natural gas from fracked shale rocks often contains little helium so Bakker's team targeted a formation in Montana similar to a conventional natural gas reservoir but filled with volcanic gases, including helium. Avanti plans to open its first plant early next year.

Gottwald is hopeful the government will grant further delays to a sale that was originally scheduled for 2021. "We have a lot of folks who share our concerns," he says. Garvey wonders whether anyone will bid on the reserve at all. Once the helium is extracted, the reservoir could be used to store other gases, but for now, the sprawling government asset is beset by maintenance issues and litigation. "We won't know until they open those bids," Garvey says. "Believe me, we are all kind of curious." ■

Julia Rosen is a science journalist in Portland, Oregon.

Buoyant prices

The U.S. government sold crude helium from a strategic reserve until a last auction in 2018. Researchers worry disruptions from a planned sale of the entire reserve system could exacerbate an ongoing helium shortage and send prices even higher.



CREDITS: (GRAPHIC) C. BICKEL/*SCIENCE*; (DATA) U.S. GEOLOGICAL SURVEY

CONSERVATION

How domestic cats wiped out the Scottish wildcat

After millennia of isolation, a few decades of interbreeding have rendered the animal "genomically extinct"

By **David Grimm**

Though it lies in ruins on the northeast coast of England, Kilton Castle was once an imposing stone fortress, home to several noble families, and—it appears—at least eight cats. Archaeological excavations in the 1960s uncovered a well, at the bottom of which lay the bones of several felines dating back to the 14th century. The animals were an odd mix: Some were domestic cats, but other, larger specimens appeared to be European wildcats, a fierce, burly species that has inhabited the continent for hundreds of thousands of years.

The two species' closeness in death was deceptive. A study published this week in *Current Biology* finds that even though European wildcats and domestic cats overlapped in Great Britain for more than 2000 years—including at sites such as Kilton—they appear to have almost never interbred. That changed suddenly about 70 years ago, when domestic cats began to mate with wildcats in Scotland. In the span of mere decades, the genome of the Scottish wildcat—the last remaining wildcat in Great Britain—has become so corrupted that the animal is now effectively extinct, a second study in the same issue finds. The findings could complicate ongoing efforts to save the most endangered mammalian carnivore in Great Britain.

"It's very intriguing work," says Shu-Jin Luo, a geneticist at Peking University who researches the DNA of wildcats in China and who has similar concerns about the impact of domestic cats there. "The studies set a very good template for studying the interactions between domestic cats and wildcats around the world."

Domestic cats have been in Europe for

thousands of years, having likely followed migrating farmers after originating in the Middle East. Once they entered the continent, they invaded the home of a distant relative, the European wildcat.

The two aren't so different: Both belong to the genus *Felis* (our kitties are *F. catus*; the wildcat is *F. silvestris*), and the European wildcat has a brown, striped coat reminiscent of some tabbies. But the wildcat is about 20% bigger, with denser fur and a thick, blunt tail. It's also fiercely solitary and wants nothing to do with humans. Try as you might to tame it, says Roo Campbell, a mammal specialist at NatureScot, Scotland's nature agency, "you're not going to end up with a cuddly pet."

So perhaps it's no surprise that domestic cats and European wildcats kept to themselves. Even though domestic cats became widespread in Europe during Roman times, DNA markers from modern cats and ancient bones show virtually no genetic overlap between the two species, the first paper reports.

One explanation is that—Kilton Castle aside—they largely avoided the same places. Domestic cats lived in close proximity to people, where food (and perhaps petting) was plentiful, while wildcats preferred scrubland far from human habitation. Wildcats also only mate twice a year, whereas domestic cats can mate whenever they want, so hooking up at the right time would have proved challenging.

Dogs and wolves show a similar pattern, notes study author Greger Larson, an evolutionary biologist at the University of Oxford. Those two canids have overlapped in the Northern Hemisphere for at least 11,000 years, he says, but show little evidence of interbreeding.

Yet the barrier between domestic cat and wildcat eventually broke down. The second new study—based on DNA from dozens of modern domestic cats and wildcat samples from nature, captivity, and museums—finds that, starting in the mid-1950s, more than 5% of the genetic markers in Scottish wildcats began to resemble those of domestic cats. After 1997, that figure jumped to as high as 74%.

That may be because the Scottish wildcat had nowhere else to go, says Campbell, an author of the study. Human hunting of wildcats—for pelts and because they were seen as vermin—in Great Britain during medieval times eventually drove the animals up into the remote highlands of Scotland. "It was their last stronghold," he says. Here, they began to recover. But in the 1950s, a viral disease decimated the rabbits the cats relied on, while human encroachment stole critical habitat. As their population shrunk—by some estimates it became as low as 30 individuals—they had no choice but to mate with domestic cats.

Captive Scottish wildcats, part of the Royal Zoological Society of Scotland's Saving Wildcats Project, which has begun to release the animal back into the wild.

Today, the genome of the Scottish wildcat is so "swamped" with domestic cat DNA that the animal is "genomically extinct," the authors conclude. All that's left in nature is a "hybrid swarm," they write, a confused mix of wild and domestic DNA. In some cases, the wildcats' striking stripes have been replaced by spots, as well as other patterns and colors not found in nature.

"Everything these wildcats have evolved over thousands of years is being lost in a few generations," says the study's lead author, Jo Howard-McCombe, a conservation geneticist at the Royal Zoological Society of Scotland (RZSS). The findings are a particular sting to Scotland, where the wildcat is a symbol of bravery and independence, appearing on everything from high school logos to the crests of storied clans.

Ironically, the one thing that may have kept Scottish wildcats hanging on so long is genes they inherited from domestic cats. The researchers found that wildcat DNA was heavily enriched in domestic "major histocompatibility complex" genes, which help our pets fight deadly diseases such as feline leukemia and feline immunodeficiency virus (similar to HIV in humans). But the only reason wildcats would have been exposed to these diseases in the first place is close contact with domestic cats, Howard-McCombe says. "It's like they're stabbing you, then calling the ambulance."

Some hope may lie in the 160 Scottish wildcats that live in captivity, thanks to early conservation efforts. Only 18% of their DNA markers register as domestic, according to the new work. Helen Senn, an author on both new papers and the leader of RZSS's Saving Wildcats Project, is spearheading an effort to reintroduce the captive cats back into the wild. This summer, her team released 19 of the animals into a national park in the Scottish highlands, far from domestic cats. The team hopes that as the wildcats adapt to their environment over several generations, they'll begin to shed their domestic DNA.

The plan sounds dubious to Steve Piper, former chairman of the Scottish Wildcat Association, a conservation charity. He believes the captive wildcats are genetically too far gone to rescue. A better solution, he says, would be to introduce wildcats from Romania or elsewhere in Europe, where they're still thriving.

Senn remains optimistic, however. She knows that it may never be possible to return to the Scottish wildcat to 100% of its genetic glory. But, she says, "We've got to start somewhere." ■

WORKFORCE

Campus unions boost pay and squeeze PIs

Since January 2022, more than 35,000 U.S. graduate students and postdocs have unionized

By **Katie Langin**

After months of tense negotiations and a barely averted strike, last week postdocs and associate researchers at Columbia University agreed to a contract that will boost their minimum salary by $10,000, to $70,000, and provide other benefits, including a $5000 child care allowance. Credit goes to their labor union, which formed in 2018, says Elsy El Khoury, a chemistry postdoc and a union organizing committee member. "We got to where we got because we were ready to strike."

It's one of the latest developments in a rising tide of campus labor organizing efforts across the country—a tide that's forcing the academic community to reckon with how it compensates early-career researchers and what the future of academic science will look like. Many graduate students, postdocs, and established scientists see the rise in union formation and strike-related activity as essential for forcing universities—which have historically opposed organizing efforts—to enact long-overdue improvements in pay and working conditions. But questions linger about where the money will come from and how universities and principal investigators (PIs) will adapt to the added labor costs. "The cost of doing business is escalating because … unions have been very successful," says Nicholas DiGiovanni, a labor and employment lawyer at the firm Morgan, Brown & Joy who represents institutions in their dealings with graduate student unions.

The surge in unionization activity has been especially pronounced among graduate students. Since January 2022, the number of students represented by graduate student unions on U.S. campuses has grown by more than 33,000, according to data compiled by National Center for the Study of Collective Bargaining in Higher Education and the Professions at Hunter College. (In the same time frame, unionized postdocs increased by roughly 2000.) Support for unionization has increased as well: During unionization votes in 2022 and the first half of this year, 91% of students were in favor of forming a union—up from 67% from 2017 to 2019.

William Herbert, executive director of the center, attributes the shift to a combination of factors, including the COVID-19 pandemic, which drove people to rethink how they want to be treated in the workplace. He also sees a generational shift away from individualism and toward collective action, citing polling data showing an "extraordinary degree of support [for labor] among people under 30."

Long-standing issues around pay and working conditions continue to play a major role as well. "Grad workers everywhere are realizing we face a lot of crises; … cost of living has gone through the roof," says Michael Mueller, a math Ph.D. student at the University of Michigan (UM), where graduate students ended a 5-month strike in August. "And our employers are not going to give us what we need unless we stand up and win it for ourselves."

The success of a high-profile strike at the University of California (UC) system added fuel to the movement. Last winter, UC graduate students and postdocs won contracts that raised their base pay by thousands of dollars, among other benefits. Such gains create a "snowball effect," inspiring early-career researchers elsewhere to form unions and push for similar agreements, says Acacia Patterson, a materials science and engineering Ph.D. student at Washington State University, where graduate students are currently negotiating a new contract. Mueller adds, "Every time we see that another union won a new victory, it sets the bar higher."

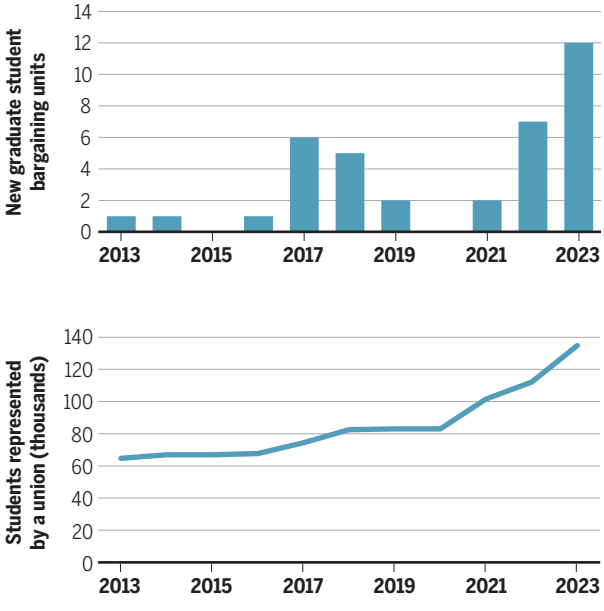
Many university administrators agree higher wages are necessary. But they also face very real budgetary constraints. Mary Sue Coleman, past president of the Association of American Universities, notes, "Administrators must balance salaries and benefits across the employee spectrum, not simply for one group." Eric Nestler, dean for academic and scientific affairs at the Icahn School of Medicine at Mount Sinai, where a union representing postdocs is currently negotiating a new contract, adds: "For people to think that … somehow there's this magic pot of money that we are all sitting on and not using, it's just not true."

Raises are easier to budget for if the yearly jumps are modest, in the "single digit percentages," says Michael Solomon, dean of UM's Rackham Graduate School. In August, his university signed a contract with the union representing graduate student instructors and teaching assistants that provides an 8% raise during the first year of the contract, and 6% annually after that. "The way that we are accommodating it is by continuing to pay really careful attention to the size of cohorts, so that we really are sure that we have the resources available," Solomon says.

Individual faculty members, who support grad student researchers and postdocs out of their own grants, are having to take a careful look at their budgets as well. "Every lab is in many ways its own little microbusiness," says Lisa García Bedolla, vice provost for graduate studies and dean

Getting organized

Over the past 2 years, the U.S. academic community has seen a marked uptick in new graduate student bargaining units (top) and in the total number of students represented by unions (bottom). (This year's data include new unions created through June; unions representing only undergraduate students are not included.)



GRAPHIC: C. BICKEL/*SCIENCE*

of the graduate division at UC Berkeley. In her view, the raises granted in the new UC contracts were necessary given the high cost of living in California, but their size and suddenness created challenges—especially for "faculty who reasonably didn't expect to have that additional cost placed on their grants," she says. "When the increases are 10%, 20%, that's just a challenge for anyone's budget to absorb."

García Bedolla says UC Berkeley gave faculty members some additional support to pay for salary increases this year. But, "We don't have a lot of new sources of funds to put towards those growing expenses," she emphasizes. In the future, she expects PIs will have to look at their funding and "decide how best to staff their labs, given their many competing needs." The senior scientists and administrators *Science* spoke with expect this will be the case at many institutions. "I think the size of the average lab will have to decrease," Nestler says.

Many union representatives acknowledge that federal research grants haven't kept up with inflation and say they worry about the extra burden on lab leaders. During contract negotiations, some unions have pushed for universities to allocate funds to pay for wages and benefits. "We said it explicitly at the bargaining table that we want central money mobilized, not to have all this—all these raises, and all this weight—carried by PIs only," El Khoury says.

At Rutgers University, where 9000 graduate students, postdocs, and faculty engaged in a weeklong strike earlier this year, the university did agree to fund raises that weren't budgeted into grants. But Columbia's union wasn't able to secure a similar deal. El Khoury says the union has been in contact with professors to say that "if they want to pressure the university and ask for help as PIs, we are going to be as a union supporting that."

Some academics say federal funding agencies such as the National Institutes of Health and the National Science Foundation need to boost grants to cover the higher pay. "The universities and those of us who work in them need to demand federal funding agencies do more," says Rebecca Givan, an associate professor of labor studies and employment relations at Rutgers and the general vice president of the union that represents academic workers at her institution.

Regardless, given rising salaries and inflation, "less research is going to be produced per dollar of [grant] money," Nestler says. Still, the need to pay graduate students and postdocs better is clear: "We have to do this. This is the way the … research enterprise will have to change." ■

ECOLOGY

Global algal outbreaks are crowding out corals

Reefs already stressed by bleaching are being smothered by crust-forming red algae

By **Elizabeth Pennisi**

About 11 years ago, Peter Edmunds noticed small dark red splotches on the reefs off the Virgin Islands that he'd been surveying for decades. The culprits, red algae that form a crust on underwater surfaces such as coral, were previously known to grow only in small patches, tucked away in crevices among the reefs. Since he first noticed them, they've "just gone gangbusters," says the marine biologist from California State University, Northridge.

And not just at his study sites. Edmunds and colleagues report this week in *Current Biology* that these algae are spreading rapidly in the Caribbean Sea and elsewhere, killing existing corals and crowding out new ones. The authors don't have a solid explanation for the algae expansion, although warming waters or another aspect of climate change may be a driver. But they and others worry this new menace will hasten the demise of ecosystems already decimated in many places by multiple bleaching events, many also linked to climate change. "They are apparently a much more serious threat to reefs than I realized," says Nancy Knowlton, an emeritus coral reef biologist at the Smithsonian Institution's National Museum of Natural History.

These latest coral killers are a group of more than 140 hard to distinguish red algal species belonging to the Peyssonneliaceae family. Some scientists mistake them for coralline algae, which also form crusts on reefs but help promote growth of the living structures. Whereas coralline algae form thin, hard crusts that are pink or whitish, peyssonnelid algae make thicker, brown or dark red crusts that are often a little squishy above a hard base. Fish love to eat the former but tend to avoid the darker algae, Edmunds says, allowing them to grow unchecked until they smother corals to death. Peyssonneliaceae also keep damaged coral from regenerating by preventing drifting coral larvae from settling and maturing into sedentary adult polyps, Edmunds says. The only place he's seen settled larvae is where sea urchins have eaten away the crust.

Edmunds is well acquainted with other algal threats to reefs, including seaweed (macroalgae) that can grow on degraded reefs and prevent them from recovering. "It's old news that corals are dying and macroalgae [are] flourishing," he says. "But this is a new hidden player that's crept up and bitten us in the butt."

After he first noticed the red algae in 2012, Edmunds tracked their spread at his study sites. By 2019, they covered between

A graduate student assesses young coral growing beside encroaching red algae (brown crust).

47% and 64% of the coral surface 3 meters down on this Caribbean reef, he and his colleagues found. He and other researchers have documented obvious swaths of these algae off Puerto Rico and in one bay in Bonaire. The algae also seem to be expanding across the Indo-Pacific, according to Tom Schils, an algae specialist at the University of Guam and a co-author of the new study.

In Hawaii, the Mariana Islands, the Marshall Islands, the South China Sea, and Vanuatu, the crusts have overgrown corals killed by coastal runoff and other environmental factors. In Guam, Schils and colleagues monitored reefs between 2016 and 2021 and found 48 peyssonnelioid species new to the area. Red algae are "proliferating and there's more diversity, but it's not clear why," says Sara Cannon, an aquatic biologist at the University of British Columbia.

"This is the kind of phenomenon that's easily overlooked if you're not paying attention," says University of Queensland marine ecologist Peter Mumby. The wide distribution of the algae in the Pacific Ocean surprised Jason Spadaro, a marine ecologist at the Mote Marine Laboratory "These algae may be a potentially new dominant organism on coral reefs globally," he says.

Others are less certain they are a big threat. Terence Hughes, director of a coral reef center at James Cook University, wants more evidence, saying there are no fresh data in the paper. And Robert Steneck, an emeritus marine biologist at the University of Maine who is among the few who has long studied crustose algae, has not seen any increases on the reefs in Bonaire, the Dominican Republic, or Belize.

Jeremy Jackson, an emeritus marine ecologist at the American Museum of Natural History and the Smithsonian Tropical Research Institute, also isn't persuaded the algae are running rampant. He wishes coral biologists had followed the lead of tropical ecologists studying forests, who long ago established rigorous and standardized procedures for tracking plots, making it possible to perceive global long-term trends. "This paper epitomizes the problem," he notes.

But, Mumby says, "It's important to draw attention to this trend and motivate a new generation of researchers to study [these red] algae more carefully." Although their impact may not be as catastrophic as coral bleaching, they could tip corals already stressed by climate change over the edge, he points out. "Corals do have yet another coral killer," Steneck agrees. "And this is truly bad news." ■

ILLUSTRATION: SMARTBOY10 AND AKEOLIMB/ISTOCK PHOTO. ADAPTED BY N. JESSUP/*SCIENCE*

POLICY

Debate brews over U.S. census changes to disability questions

The proposed changes would decrease the percentage of people considered disabled, affecting data and resources

By **Phie Jacobs**

The U.S. Census Bureau may soon change the way one of its nationwide surveys asks about disability. But alarm bells are ringing for many researchers and activists, because the proposed change would dramatically decrease the official number of people in the U.S. who are considered disabled. "Disabled people are already underserved," says Scott Landes, a sociologist at Syracuse University who studies disability. Altering how the Census Bureau gathers disability statistics will generate "inaccurate information," he says. In an 18 October letter, he and other disability researchers and advocates asked the Census Bureau to reconsider.

The change concerns a section of the bureau's annual American Community Survey (ACS), which is an important source of demographic, social, and economic data. The ACS uses a set of six yes-or-no questions—related to difficulty with hearing, vision, and other functions—to determine disability status. A "yes" answer to any of those questions means the respondent is disabled. Many state and federal programs allocate funding based on ACS data, which are used to evaluate whether disabled people have equal access to housing, education, and health care. But the ACS is also a vital resource for researchers. "It [can] drown out a lot of better designed sources of disability data," says Jaime Seltzer, a disability activist and researcher at Stanford University who uses the ACS for her work on chronic fatigue syndrome. These days, Seltzer says, many are also using the data for Long Covid research.

Following a September meeting with the Census Scientific Advisory Committee (CSAC), bureau officials are now recommending replacing those questions with a set developed by the Washington Group on Disability Statistics, a United Nations–convened organization that creates disability data collection tools for censuses and national surveys. Its method, known as the Washington Group Short Set on Functioning (WG-SS), also consists of six questions on the same topics as the ACS questions. Instead of answering "yes" or "no," however, respondents rate their level of difficulty on each function by choosing one of four options, from "no difficulty" to "cannot do at all."

Compared with the current questions, the WG-SS will "capture information in a manner that reflects advances in the measurement of disability," an agency spokesperson said in an emailed statement. But it would also reduce the prevalence of disabled people in the U.S. to 8%, down from the ACS's current estimate of 14%, bureau statisticians reported at the CSAC meeting. This decrease is because WG protocol states only those who indicate "a lot of difficulty" or "cannot do at all" for one or more of the questions should be categorized as disabled.

Jennifer Madans, who serves as WG's chair, says this cutoff was chosen because it yielded the most consistent responses when the WG-SS was tested internationally. In that context, "You want everyone to be using the same definition of disability," she says, adding, however, that individual countries should use the cutoff that will best inform their policy decisions.

Bonnielin Swenor, director of the Johns Hopkins University Disability Health Research Center, says the change would have "massive consequences for a group that is already struggling, pleading, begging for our data to be collected in a more robust way." Swenor, who was the lead author of the 18 October letter, also says asking respondents to rank their difficulty level "entrenches a hierarchical view of disability"

that implies some are less deserving of accommodations.

Madans says the scaled response system better captures the "continuum" of disability. The yes-or-no format of the ACS questions, she argues, may obscure crucial differences in functionality, making it seem as though people with significant functional difficulties are doing better than they really are.

Neither method is perfect. Both question sets do a poor job of accounting for people with psychiatric and chronic illnesses, says Jean Hall, director of the Institute for Health and Disability Policy Studies at the University of Kansas. To alleviate this gap, Hall and her colleagues developed the National Survey on Health and Disability (NSHD), which allows respondents to describe and self-categorize their disability. Hall and her colleagues recently reported that the ACS questions didn't identify nearly 20% of people who reported disabilities to the NSHD, and the WG-SS missed 43%. Switching to the WG-SS would mean "going from a measure that's not great to one that's much worse," Hall says.

Swenor and others also claim the bureau didn't discuss the change to the questionnaire with the disability and disability research communities before proposing it. Bureau officials told *Science* the agency "regularly engages with disability researchers and data users," but the October letter alleges that the bureau's plan was to engage with the disabled community after the change was made. "This approach to community engagement is not acceptable," the letter states, while also calling for the creation of a national task force to develop better methods for gathering disability data.

Daniel Mont, who serves as chair of WG's Analytical Working Group, says that although the bureau's proposal may have left some U.S. researchers feeling blindsided, the WG-SS itself has undergone extensive testing internationally. He also says the bureau could report two different numbers: one representing the prevalence of Americans who face significant difficulties—the 8% figure the bureau calculated—and a statistic that captures more of the disabled community. Bureau statisticians found that defining disability to include those who reported "some difficulty" put its prevalence at almost 32%.

The proposed change will be open for public comment until 19 December, although the National Advisory Committee, which advises the Census Bureau on policy and research issues, will discuss initial comments on 16 November. Landes hopes policymakers will listen to the voices of disabled Americans and not move forward with the change. "The power of the disability community is strong," he says. ■

SCIENTIFIC COMMUNITY

Senate confirms Monica Bertagnolli as NIH director

After 2-year gap, biomedical research agency has a permanent chief, but she faces many challenges

By **Jocelyn Kaiser**

At last, the world's largest biomedical research agency has a permanent leader. The U.S. Senate this week voted 62-36 to confirm oncologist Monica Bertagnolli to direct the U.S. National Institutes of Health (NIH).

The 7 November confirmation of Bertagnolli, who is replacing acting Director Lawrence Tabak, brought relief to the biomedical science community. Researchers had been waiting for a new NIH chief since longtime director and geneticist Francis Collins left in December 2021. President Joe Biden nominated Bertagnolli in May.

A cancer surgeon and clinical trial expert, Bertagnolli is NIH's 17th director and the second woman to lead the agency. She left Harvard Medical School and affiliated hospitals to head the National Cancer Institute in October 2022. She now takes the reins of the $47 billion NIH as it faces enormous budget and political pressure and with just 15 months remaining until the end of Biden's term—and perhaps her tenure. "She will have plenty of problems on her plate," says cancer researcher Harold Varmus, who led NIH in the 1990s.

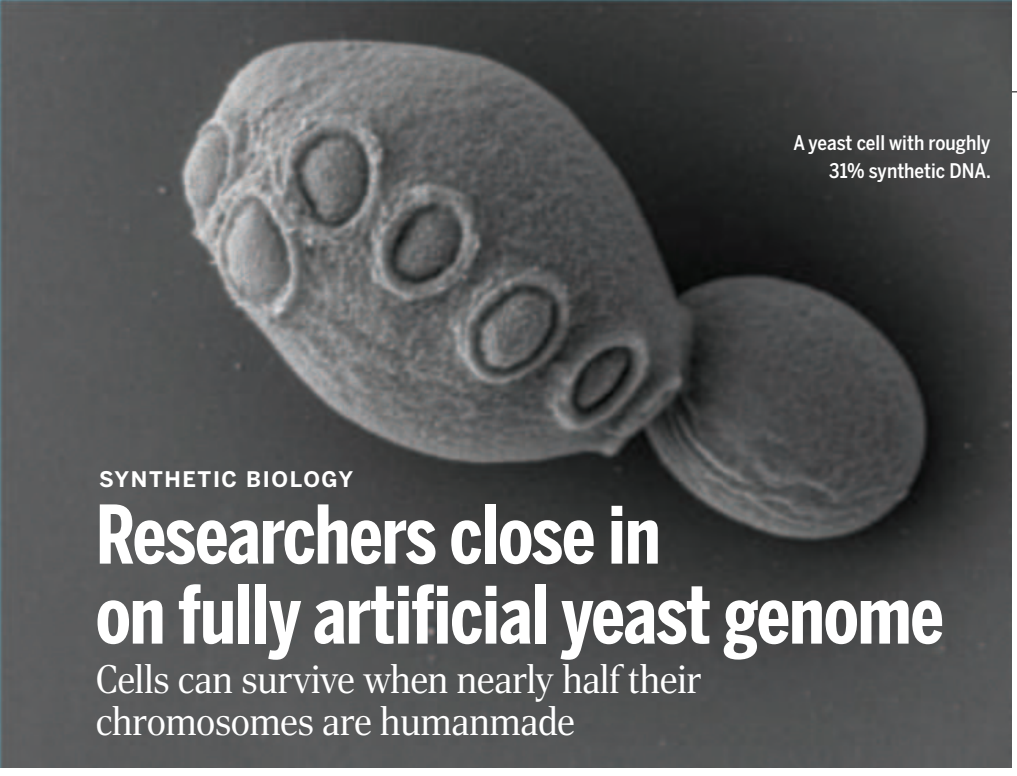
The substantial number of votes against Bertagnolli, for example, reflects concerns among both conservative and liberal lawmakers about NIH policies. (In contrast, the Senate unanimously confirmed Varmus and Collins.) Conservative lawmakers have complained about NIH's role in funding politically sensitive research such as gender-affirming care and virus studies that some claim led to the COVID-19 pandemic. Two liberal senators, John Fetterman (D–PA) and Bernie Sanders (I–VT), chair of the Senate's health committee, voted against her confirmation because of concerns that NIH is not doing enough to bring down drug prices.

At her confirmation hearing before the committee last month, Bertagnolli said she wants to bring NIH's innovations and clinical trials to diverse populations in the United States. One immediate challenge she will face is urging Congress to preserve NIH funding as it works to reconcile a 2024 spending measure passed by the Senate that would slightly increase the agency's budget with a bill from the House of Representatives calling for deep cuts. "She's going to have to fight for the least amount of reduction or even a flat budget or maybe a 1% increase," says former Director Elias Zerhouni, who led NIH in the mid-2000s. Even with a stagnant budget, NIH has committed to funding a hefty number of young investigators, which means mid- and late-career investigators could face a disproportionate budget squeeze.

Building support among skeptical lawmakers will be a priority. Republicans in the House, for example, have used the 2024 spending bill and an accompanying report released last week to highlight some of their concerns. The measure would defund NIH efforts to promote a diverse biomedical workforce, study gender-affirming hormone treatment, set up an office on sexual and minority health research, and support fetal tissue studies. The final bill negotiated with the Senate is not expected to include those funding bans, notes Jennifer Zeitzer of the Federation of American Societies for Experimental Biology, but Bertagnolli is likely to face ongoing questions about those efforts and potentially risky pathogen research.

Researchers hope Bertagnolli will also address other matters, including reversing NIH's shift away from basic research toward more applied studies. Also awaiting Bertagnolli is an advisory panel's report, due in December, on low salaries for postdocs and the increasing difficulty senior scientists face in recruiting that pivotal workforce. And she is expected to be pressed to clarify NIH's relationship with the Advanced Research Projects Agency for Health, a new independent agency within NIH that Biden created to boost high-risk, high-impact research.

It's a lot to pack into a tenure that might not last beyond Biden's term if he fails to win reelection. "It's a complicated agency" with "a steep learning curve," says Carrie Wolinetz, a former top NIH official now with the lobbying firm Lewis-Burke Associates. "I admire the fact that she's willing to take it on." ■



A yeast cell with roughly 31% synthetic DNA.

SYNTHETIC BIOLOGY

Researchers close in on fully artificial yeast genome

Cells can survive when nearly half their chromosomes are humanmade

By **Mitch Leslie**

A 17-year project to craft a synthetic genome for yeast cells has reached a watershed. Researchers revealed this week in 10 new papers that they have created designer versions of all yeast chromosomes and incorporated almost half of them into cells that can survive and reproduce. "It's a milestone we have been working on for a long time," says geneticist Jef Boeke of NYU Langone Health, director of the project.

"It's a very impressive body of work," says synthetic biologist Sanjay Vashee of the J. Craig Venter Institute, who wasn't connected to any of the studies.

Researchers have tinkered with the genomes of yeast and many other organisms using editing technologies such as CRISPR. But building a new version from the ground up opens the way to making bigger changes to an organism's genome and delving deeper into its organization, function, and evolution.

Developing synthetic genomes may also make it easier to upgrade the many organisms that are crucial for industry, agriculture, and medicine. Yeast, for example, not only keeps breweries and wineries in business, but also churns out a variety of chemicals and drugs, including insulin. With a synthetic genome, "we can more rapidly improve strains and find the genes that are important for increasing production," Boeke says.

So far, scientists have generated artificial genomes for several viruses and bacteria, but yeast would be the first eukaryote—the group of organisms, such as flies, snakes, and humans, whose cells have a nucleus. The yeast species the researchers used, *Saccharomyces cerevisiae*, carries 16 chromosomes and is a much bigger challenge than bacteria, which typically have a single chromosome and a fraction as much DNA. Launched in 2006, the Synthetic Yeast Genome Project enlisted scientists at more than a dozen institutions worldwide to tackle the problem, along with hundreds of undergraduates who helped synthesize bits of the genome.

The researchers didn't attempt to redesign the genome one nucleotide at a time. Instead, they revised the native yeast genome, adding thousands of modifications that simplify its structure, boost its stability, and make it easier to study. For instance, they carved out the transposons, itinerant stretches of DNA that can leap from location to location in the genome, disrupting DNA sequences.

They also pruned the genome by excising many of the introns, segments of DNA that don't code for portions of proteins. And to make the new yeast genome easier to manipulate in future experiments, the team included several hundred short DNA sequences that can prompt sections of chromosomes to rearrange.

In most cases, the researchers left genes on their original chromosomes. But a team led by synthetic biologist Yizhi "Patrick" Cai of the University of Manchester, international director of the project, created a new, 17th chromosome to house yeast's 275 tRNA genes. They code for RNA molecules that transport amino acids, the building blocks of proteins.

Although tRNAs are essential for protein synthesis, their genes "are a lot of trouble for the genome," Cai says, because "they are DNA damage hotspots" that can cause breaks. By isolating these disruptive genes on one chromosome, the researchers hoped to tame them. They found that yeast cells could survive and grow—albeit more slowly than unmodified cells—with this newfangled chromosome, they report in *Cell*. Synthetic biologist Paul Freemont of Imperial College London calls this work "a tour de force."

Separate teams assembled each synthetic chromosome in a different strain of yeast. "We show all chromosomes can be successfully constructed under the same design principles from scratch," says biologist Yue Shen of BGI Research, who heads one team. But the researchers needed to incorporate the chromosomes into the same yeast cell and determine whether it was healthy. Boeke and colleagues repeatedly mated cells harboring different synthetic chromosomes, eventually producing yeast that contained six full-size synthetic chromosomes and a fragment of another, but not the extra tRNA chromosome.

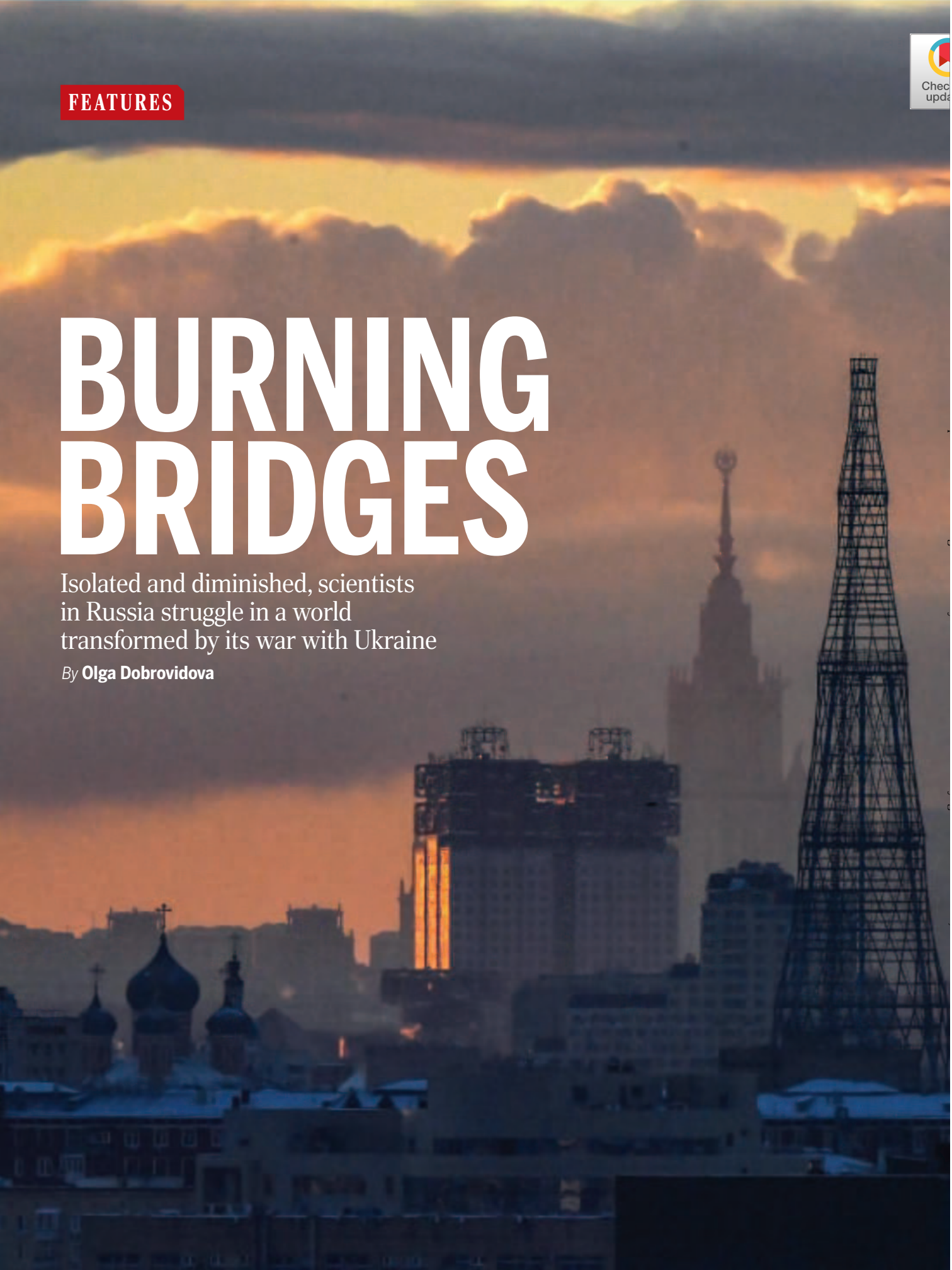
This yeast grew slower than normal because of some harmful genomic glitches, but after the researchers identified and corrected them, the strain grew about as fast as unaltered cells, the team reported in a second *Cell* paper. They then used a similar mating approach to add another synthetic chromosome, bringing the total to 7.5. In these cells, more than 50% of the DNA is synthetic. "We are more than halfway there," Boeke says.

In other papers published this week, scientists with the project used cells with the remade chromosomes to test the effects of specific structural changes. For example, moving the centromere, the junction between chromosome arms that is essential for cell division, revealed that a small stretch of DNA adjacent to the centromere helps keep chromosomes stable, Boeke and colleagues reported in *Cell Genomics*. And repositioning chromosomes within the nucleus showed that confining them near its boundary led to gene silencing, he and his co-workers revealed in *Molecular Cell*.

Synthetic genomicist Ian Ehrenreich of the University of Southern California says the results offer "a cool system for discovering rules for how DNA elements combine to produce a functional genome." Researchers are now "moving past a descriptive form of biology, to one where we understand life by building it," he says.

The team is now working to integrate the remaining chromosomes into a yeast cell and correct any genomic problems that arise. Boeke expects a yeast with a fully synthetic genome to debut in about a year. ■

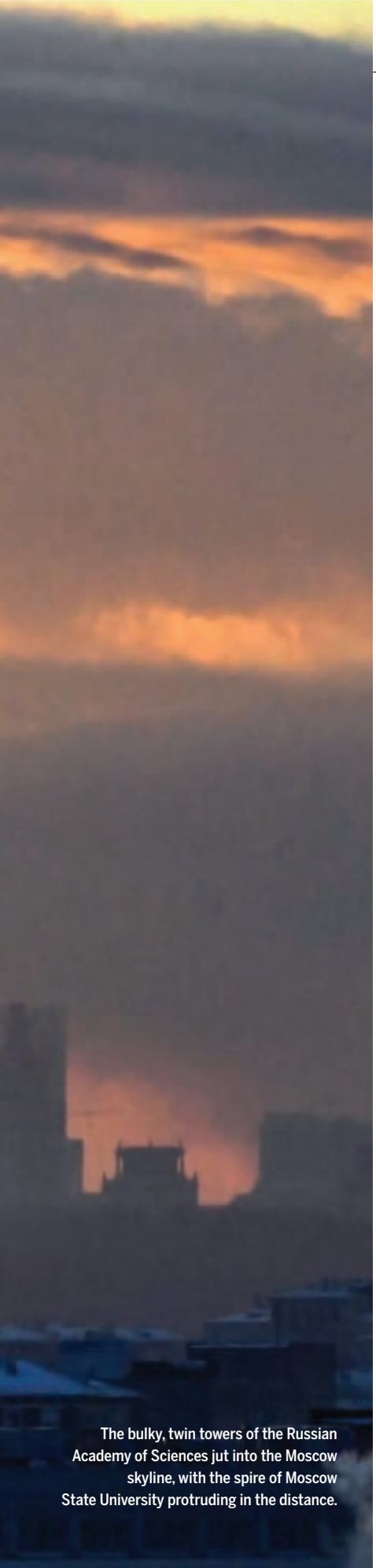
PHOTO: ZHAO *ET AL.*/*CELL*



BURNING BRIDGES

Isolated and diminished, scientists in Russia struggle in a world transformed by its war with Ukraine

By **Olga Dobrovidova**



The bulky, twin towers of the Russian Academy of Sciences jut into the Moscow skyline, with the spire of Moscow State University protruding in the distance.

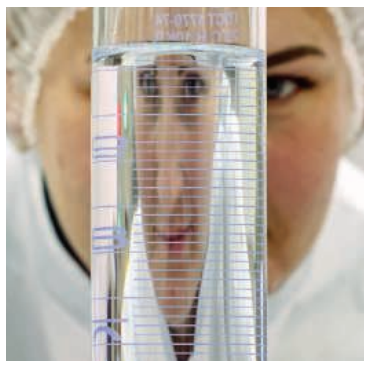
PHOTOS: (LEFT TO RIGHT) EVGENY BIYATOV/SPUTNIK VIA AP; SERGEI KARPUKHIN/TASS VIA ZUMA PRESS

Yuri Kovalev remembers how some of his older colleagues took offense when he moved to the United States in 2003 to take a postdoc position at the National Radio Astronomy Observatory. He could have stayed in Russia where he got his Ph.D., at the Lebedev Physical Institute, one of the country's oldest and most prestigious science centers. Why would he ever want to leave? "But for me it felt really natural as we were now part of a bigger world," he says.

Yet Kovalev returned a few years later, drawn by the power and potential of science in his home country. He rejoined Lebedev to work on RadioAstron, an international project that linked radio dishes in the U.S. and elsewhere with an orbiting Russian satellite to create a giant virtual telescope. With the system, Kovalev and his colleagues created some of astronomy's highest resolution images, including ultrasharp pictures of the jets shot out from supermassive black holes at the center of galaxies.

Kovalev was also encouraged by growing Russian R&D spending, which at about 1.2% of gross domestic product was half of rich-world levels but far higher than during the tumultuous 1990s. Russian scientists were retooling their labs, publishing in high-ranking journals, and adapting as science funding shifted toward competitive, merit-based grants. And Lebedev was a great place to be. It had a successful partnership with CERN, the European laboratory for particle physics near Geneva, as well as a high-temperature superconductivity center of its own, dreamed up by Vitaly Ginzburg, one of the institute's seven Nobel Prize winners. "We had both freedom of movement and license to pursue any direction in research," Kovalev says.

All that has changed. The political climate in Russia began to sour in the 2010s, with President Vladimir Putin gradually cracking down on civil society. For many scientists, the 2022 invasion of Ukraine was the last straw. Many have fled to the West for political and personal reasons, and those who remain are contending with sanctions affecting the supply of laboratory basics. International collaborations have shriveled, and no new ones are in sight, says Kovalev, who moved away again in 2022, this time to Germany. "That is an enormous problem, and we are not really feeling its pinch yet," he says.



Sanctions have targeted laboratory supplies but researchers have found workarounds.

The Kremlin's crackdown on dissent and free speech makes it risky to speak out about science's troubles, and several researchers in Russia contacted by *Science* declined to be interviewed. One economist who asked for anonymity says expatriates should not criticize those in Russia who don't openly speak out against the war or the government. "They are blaming us for staying quiet—well, be my guest, come back here and tell us everything," the economist says.

And for some researchers in Russia, discussing their plight feels inappropriate amid the death and destruction that Russia has wrought in Ukraine and on its science community (*Science*, 9 December 2022, p. 1036). "It's as if you're going through this case of food poisoning and someone else nearby is having a heart attack," says Mikhail Gelfand, a bioinformatics researcher at the Kharkevich Institute for Information Transmission Problems and vice president for biomedical research at the Skolkovo Institute of Science and Technology (Skoltech).

For those who stay quiet, life in Russia can carry on mostly as normal. That's why many Ukrainian scientists want to see stronger sanctions against Russian research institutions. They say some of those institutions aid and abet Russia's military-industrial complex. "There is no doubt that blocking access to scientific equipment, international funding, high-profile collaborations, databases, and publications will make Russian science weaker, and thus, to a degree, will decrease the capacity of Russia to invade its neighbors," Yaroslaw Bazaliy, a Ukrainian condensed matter physicist at the University of South Carolina, and his colleagues said in a statement to *Science*.

Some Russian researchers aren't ready to give up on their country yet—in part because of the scientific riches it holds. For example, Russia controls a large fraction of the Arctic—a region that contains clues to the pace of climate change and its consequences. Alexander Kirdyanov, a tree-ring researcher at the Sukachev Institute of Forest, has kept up summer field trips to Siberia and plans to return to Russia once he concludes a visiting position in the United Kingdom. "If you think you can precisely study everything from space, good luck measuring the age of a tree or the thickness of moss cover by satellite," he says.

At a quantum research forum in July, Russian President Vladimir Putin (right) railed against what he called Western attacks on the nation's technological sovereignty.

Kirdyanov also feels an obligation to continue to train young researchers in Russia. Kovalev worries that, as more and more bright, young colleagues leave, maintaining a critical mass of researchers able to do high-quality science will become extremely difficult. "So, we should be beyond grateful to those top researchers who choose to stay in Russia," he adds.

As the war in Ukraine grinds on, there is little appetite for collaboration with Russia, and little hope for peace and rapprochement any time soon. But Dmitri Kuvalin, a member of the Russian National Security Council's scientific advisory board who studies Russian industry at the Institute for Economic Forecasting, still sees science as a bridge. It "connects countries now and will help reconnect in the future," he says.

JUST BEFORE TANKS rolled toward Kyiv, Ukraine, in February 2022, Irina Dezhina, a researcher at the Gaidar Institute for Economic Policy, and Elizabeth Wood, a historian at the Massachusetts Institute of Technology (MIT), published a study on 3 decades of U.S.-Russian partnerships in science. They identified numerous cases where researchers in both countries benefited from collaboration, despite geopolitical tensions and major differences in scientific cultures. They found U.S. researchers often marveled at Russian ingenuity and strong theoretical instincts, whereas the Russians adopted U.S. habits in writing good grant proposals and publishing internationally.

The timing was ironic. "We had so many plans to continue this work," Dezhina recalls. Yet at the end of that month, just 4 days into the invasion, MIT pulled the plug on its decadelong cooperation with Skoltech, an English-language university it had helped launch, and ended the joint seed grant program that had funded Dezhina's work.

A cascade of similar decisions followed. Like many other nations, Germany ended all research collaborations with Russia. That forced the Max Planck Institute for Extraterrestrial Physics, for example, to shut down one of the two instruments onboard the Russian Spektr-RG spacecraft, which was mapping the universe in x-ray wavelengths. CERN said it will cut ties with Russia when its contract expires in 2024, ending a relationship with more than 1000 Russian scientists. The U.S. hit Russian research organizations, including Lebedev, with harsh sanctions that drove away suppliers, partners, and even some employees.

Outspoken Russian scientists began to protest the war, with the first open letter published hours after the invasion began. Soon after, young researchers and university faculty mobilized to publish dozens of other protest letters.

The state hit back, detaining people at street protests and pressuring the scientific establishment. Leaders at the Russian Academy of Sciences (RAS) chastised members who had signed letters for "insulting the government." Rectors of major universities published a pro-war statement that called for supporting the army and Putin. The science community newspaper that started the initial campaign, *Troitsky Variant*, was designated a "foreign agent," a label that scared off donors and forced its nonprofit publisher to close. After the government introduced novel criminal offenses—including calling the events in Ukraine a "war"—organizers had to hide the lists of signatures.

An Orthodox priest blesses Russian conscripts called to war. Ukrainians say supporting Russian science aids and abets its military-industrial complex.

PHOTO: ALEXEY PAVLISHAK/REUTERS

"You can't watch everyone in the country, but everyone who signed those letters is certainly being watched," one signatory says. They told *Science* they had several increasingly uncomfortable conversations with their superiors about their actions later in the year. Those discussions were invariably prompted by "reports" by someone outside the organization that were never shown or discussed in detail.

Many researchers now fear being reported to the authorities by "concerned citizens" or even their colleagues. Those fears are not unfounded. Alexandra Arkhipova, a social anthropologist formerly with the Presidential Academy of National Economy and Public Administration, a civil service and management university in Moscow, has been tracking court sentences for "discrediting the Russian army" since the novel offense was created. Her team has found more than 7200 cases so far, including hundreds where offenders said something in a public setting—and were later snitched on by witnesses.

She says there is no clear pattern in who gets persecuted: "This wave of repression is expressly random." To her, that is the point; a climate of fear is more useful to the state than any sort of logic. Arkhipova, who was informed on several times herself, left Russia in 2022 and continues her work at France's School of Advanced Studies in the Social Sciences.

FOR THOSE WHO just want to carry on quietly with their studies, sanctions are making it challenging. The fallout varies by scientific field: A 2022 industry survey of more than 4000 scientists, run by the RAS Institute of Psychology and the newspaper *Nezavisimaya Gazeta*, showed some 70% of those in natural and medical sciences expected major disruptions. Some of the new hurdles are mundane, such as being unable to pay to publish articles in Western journals because Russian banks have been banned from SWIFT, an international financial transaction processing system.

But others are more technical. Russia remains far from self-sufficient in research supplies and equipment. Speaking at a conference in December 2022, Dmitry Livanov, rector of the Moscow Institute of Physics and Technology and former government minister for science, said some 80% of the Russian science market belonged to foreign suppliers—in his words, an unfortunate situation "of our own making." The country imports not only high-end devices such as DNA sequencers, but also basic equipment such as laboratory balances.

Faced with shortages and supply disruptions, Russian researchers are getting creative. On Telegram, a popular messaging app, researchers chat to trade consumables, find reviews for a lab animal vendor, or even pick up something that another lab is throwing out. Our Lab, a project supported by the government, is coordinating a database of about 500 local suppliers for items such as lab benches, infrared lasers, and petri dishes. Researchers have made their own silica gel, a staple in chromatography machines, which separate chemical compounds from mixtures. Others returned to a practice of smuggling plasmids, circular molecules of DNA used to manipulate genes, from abroad after Addgene, the U.S. nonprofit plasmid repository and distributor, stopped shipping to Russia.

In addition to facing material shortages, Russian scientists have also been isolated.

Researchers prepare to deploy buoys below the frozen surface of Siberia's Lake Baikal as part of a telescope to detect neutrinos from space.

Many conferences place restrictions on researchers with Russian affiliations—or ban them outright. This is bad for science overall, which is international by default, Kovalev says. "Astrophysics in particular can't be confined to national borders."

Some Russian researchers have attended conferences without an institutional affiliation. But Kovalev says that approach exposes the scientists to administrative risks and even legal liability in Russia for misappropriation of funds used to cover the costs of attendance. It's also an ethical problem, because researchers are expected to acknowledge the institutions that support their work.

Some in the West worry about what is being lost. Siberia, for example, is a hot spot of Arctic field research, providing valuable data on environmental change. Western research institutions, which threw a lifeline to many Russian permafrost and climate projects in the 1990s, now find themselves cut off from about half of the Arctic as collaborations wither.

Ulf Büntgen, a geographer at the University of Cambridge, has worked in the Siberian north for decades, gathering tree ring data for paleoclimatic studies in a region still poorly covered by measurements. He fears disruptions in fieldwork may break continuous observational records that took a lot of effort to build. Büntgen wants climate scientists to set aside politics and resume collaborations with Russian scholars and institutions, as he believes their data and expertise are too critical to forgo. "To tackle climate change, we have to work in Russia, with Russians," he says.

> **"To tackle climate change, we have to work in Russia, with Russians."**
>
> **Ulf Büntgen**, University of Cambridge

Bazaliy and his colleagues reject that idea. Setting politics aside will only help the Russian government return to "business as usual," they say. "You might think you are helping scientists in Russia, but in reality, you may be assisting Mr. Putin."

PERHAPS THE MOST lasting damage the war has done to science in Russia is to accelerate the ongoing exodus of scientists. A sensitive subject since the 1990s, its magnitude is hard to gauge. But Johannes Wachs, who studies computer science at Corvinus University of Budapest, says emigration among the tech community can give a sense of the potential scale. He analyzed GitHub, a popular open-source developer site, for changed or deleted location information in its list of software developers. He estimates that between 11% and 28% of Russia's developers have left since the war began.

Another clue comes from the 2022 industry survey, which asked scientists how the "special military operation" affected their intention to leave Russia. One-third of respondents said it "somewhat" or "strongly" increased their intention to leave. For scientists under age 39, that figure was slightly over 50%.

In lieu of hard numbers are personal stories. Ilya Schurov, a mathematician who left Russia in early March 2022 after protesting the war, found a 2-year postdoc position in condensed matter physics at Radboud University in the Netherlands. His former employer, the National Research University Higher School of Economics (HSE), has lost about 700 faculty since the war began, its co-founder Andrei Yakovlev said in a Facebook post announcing his own departure in

August. Schurov says it was a big step down to trade a senior position for a postdoc, but he appreciates his academic freedom. "At least I can be sure I won't get fired at any moment if someone doesn't like something," he says.

That's what happened to Dinara Gagarina, a digital humanities researcher at a local branch of HSE. She was abruptly dismissed from all her projects and later fired for "amoral behavior" after she made antiwar posts on social media. Gagarina fought her termination because she wanted to keep supervising students who were preparing to graduate. Early this year she, too, left Russia after receiving calls from the police. She ultimately landed a job at the University of Erlangen-Nuremberg in Germany. "I think I've been unfairly fired, and that was really hard for me," Gagarina says. In June, a local court rejected her second appeal.

A decade ago, Dmitrii Musolin, an entomologist who studies forest pests, had returned to Russia after 10 years of research in Japan because he saw "positive signals" of change. He joined the Saint Petersburg State Forest Technical University and eventually became its vice rector, overseeing research and international affairs. "But in reality, it all went downhill from there," he says. In March 2022, he left Russia for a job at the European and Mediterranean Plant Protection Organization, just as his university boss in St. Petersburg signed the prowar declaration.

Unlike those who left the ruble of the Soviet Union, today's scientists are better equipped to find new homes for themselves abroad, in part thanks to all those international partnerships. And they join a big and diverse community abroad ready to help both Ukrainian and Russian researchers, often with little regard for passports.

"I feel equally close to both groups, I grew up in the Soviet Union after all," says Alexander Kabanov, CEO of the Russian American Science Association, a nonprofit hub for the Russian-speaking research diaspora that recently launched a mentorship network for scholars at risk. Yet he feels those fleeing bombs take priority: Kabanov has supported a Ukrainian pharmaceutical scientist who joined his nanomedicine lab at the University of North Carolina, Chapel Hill, and he has helped several others find work.

Dmitry Rudenkin, a sociologist and nonresident fellow of the Russia Program at George Washington University, finds the current wave of migrants different from earlier ones. Back then, people were on a quest for a better life. Now, they are driven by fear. "They were leaving *for* somewhere else," he says, "and now we are only running *from* Russia." Like many researchers studying this exodus, he is also part of it: When the invasion started, Rudenkin was on a work trip abroad and decided not to return home to a job at Ural Federal University.

Most exiles are not rushing to cut ties with Russia. Two senior researchers abroad, who spoke to *Science* anonymously to avoid professional repercussions, have maintained their affiliations at their home research institutions, although neither is getting paid. They say it's a gesture to co-workers that they are not abandoning ship, and a way to keep helping junior colleagues. "Why should I be the one to quit? If they want me gone, they will have to fire me," one of the researchers says.

Russia holds most of the world's permafrost. A researcher in Yakutsk cuts a sample.

PHOTO: MLADEN ANTONOV/AFP VIA GETTY IMAGES

IN JULY, Putin paid a visit to a swanky quantum research forum in Moscow. In a speech, he said that by limiting Russia's access to technology, "ruling elites" of some countries wanted to pressure Russia into giving up its sovereignty. "Russia will only go forward and follow its own road but without isolating itself from anyone at the same time," he said.

State propaganda has amplified the claim that Russia welcomes collaboration but is being targeted by unprovoked and unprecedented sanctions. "Destroying Russian science" is one of the top priorities for the West, Security Council Secretary Nikolai Patrushev told the council's scientific advisory board in April.

The government has turned to "friendly" states, bolstering joint Russian Science Foundation funding schemes with China and India and launching a new one with Iran. But many of these countries have scant, if any, history of large-scale research cooperation with Russia.

To entice researchers from overseas, the government revamped its megagrant program, which, like China's Thousand Talents Program, is intended to lure expats to Russian institutes with generous lab support. Kabanov, who received one of the first grants in 2010, believes few, if any, leading scientists working in the West would be tempted to join now.

Despite the grim mood, none of the Russian researchers interviewed by *Science* expects the massive Russian scientific sector to wither quickly. The slow-moving nature of many projects, with budgets set years ago, has also insulated some scientists from the war and its impacts, creating a patina of normalcy. An attendee at a Moscow science policy conference in April was stunned by an atmosphere they described as delusional optimism—"It's all great and going to get better."

But Gelfand applauds Russians who are continuing their work, even under the pressures of an authoritarian state. He also sees some hope in those who keep doing science in exile. "The tree may be dead, but you can graft an offshoot elsewhere, and you can even bring it back to the same place later," he says.

Kabanov, who moved to the U.S. in 1994, spent the past 2 decades helping rebuild science in his native country, even as he saw Putin consolidating power and cracking down on civil society. "Even though I realized what was going on, I was hoping for the best, just as the whole global research community did," he says.

But Musolin, who gave his homeland a second chance when he returned from Japan, doesn't see himself giving Russia a third, at least not anytime soon. "I don't think change will happen in my lifetime," he says, with little pause. ■

Olga Dobrovidova is a science journalist in Paris who does climate communications work. In 2020–21, she worked as a copywriter at Skoltech.

PERSPECTIVES

PHYSICS

Circling back to magnetism

Ultrafast experiments unveil control of magnetization with atomic rotations

By **Robert A. Kaindl**

In the classic Einstein–de Haas effect (*1*), a ferromagnet is set into rotation upon switching its magnetization, yielding insight into the microscopic nature of magnetism by evidencing the transfer of angular momentum from electron spins to the macroscopic crystal. Even now, the quantum nature of coupling between the electronic angular momentum and atomic crystalline structure remains intriguing. Excitation of ferromagnets with light revealed a surprisingly fast, femtosecond loss of magnetization, provoking extensive efforts to clarify how angular momentum is transferred so quickly out of the spin system (*2, 3*). Ultrafast diffraction studies indicate a transfer to microscopic crystal strain and to rotating motions of atoms around their equilibrium positions, called chiral phonons (*4, 5*). On page 698 of this issue, Luo *et al.* (*6*) flip the physics around to reveal how such chiral phonons can polarize the spins of a rare-earth paramagnet, akin to the application of a tesla-strength magnetic field.

Chiral phonons represent distinct crystal lattice excitations, because their rotational atomic motions entail an angular momentum and breaking of time-reversal symmetry (TRS) (*7*). TRS requires the physics to remain unchanged when time is reversed, which is violated by chiral phonons through their clockwise and counterclockwise polarizations. Breaking TRS affects selection rules—principles that determine allowed or forbidden transitions—in scattering processes and light–matter interactions. Chiral phonons can either occur as natural stand-alone modes—for example, in particular two-dimensional (2D) dichalcogenides (*7, 8*) or bulk crystals (*9*)—or they can be selectively driven by circularly polarized light, whose electric field vector rotates with time. Circularly polarized light can superimpose the motions of orthogonal lattice vibrations, resulting in unidirectional rotations of atoms around their equilibrium positions. As recently discovered carriers of broken TRS, chiral phonons are attracting extensive experimental and theoretical research exploiting their properties in semiconducting, photonic, topological, and quantum materials (*7–10*), as well as for fundamental investigations and applications of magnetism.

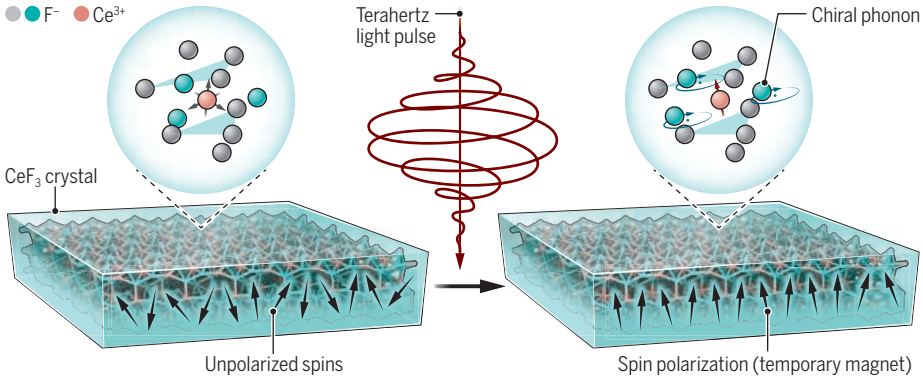
Luo *et al.* exposed a crystal of a rare-earth compound, cerium trifluoride (CeF_3), to intense circularly polarized terahertz (THz) light pulses. In its ground state, CeF_3 is a paramagnet—a material whose spins lack magnetic ordering in the absence of a magnetic field. When tuned resonant to orthogonal vibrations at 10.5 THz, the THz pulse drives into rotation predominantly the three in-plane fluoride (F^-) ions that surround each central cerium (III) (Ce^{3+}) ion (see the figure). In turn, these periodic atomic displacements dynamically modulate the crystal field, that is, the combined electric field of the F^- ions at the Ce^{3+} cation, which determines the splitting of its electronic orbitals. In search of magnetic polarization, a second femtosecond optical pulse is sent through the excited region as a probe of Faraday rotation—a phenomenon in which the polarization of light is rotated by a magnetized medium (*3*). These measurements reveal a rapid buildup of transient magnetization, which lasts for several tens of picoseconds (10^{-12} s) and thus well beyond the THz driving pulse.

To distinguish whether the observed transient ferromagnetic alignment of the Ce^{3+} 4*f*-electron spins results from chiral phonons or instead arises from direct THz excitations between electronic crystal field levels, Luo *et al.* used an optical second-harmonic generation probe. This requires breaking of inversion symmetry, either from an additional electric field or suitable distortions of the CeF_3 crystal lattice. Because the F^- anions are asymmetrically displaced from the central Ce^{3+} during chiral phonon motion, inversion symmetry is broken, and the second-harmonic probe indirectly detects these movements. The helicity—direction of rotation—of the second-harmonic signals indicates the chirality of the phonon displacements and, as the experiments show, is maximized alongside the Faraday rotation signal that is sensitive to magnetization. This correspondence demonstrates that the transient magnetization indeed arises from the THz-driven chiral phonons.

These observations are part of a rising effort to explore the microscopic coupling of spins and broken TRS lattice motions. Other studies include seminal experiments using mid-infrared pulses to excite the rare-earth antiferromagnet erbium ferrite ($ErFeO_3$) (whose neighboring spins are antiparallel, except for a small out-of-plane tilt), which set into motion elliptically polarized atomic rotations evoking spin precession analogously to a magnetic field (*11*). Moreover, two recent preliminary studies reported achieving

Chiral phonons mediate between light and magnetism

Paramagnetic cerium trifluoride (CeF_3) is excited with circularly polarized terahertz light pulses resonant to infrared active crystal lattice vibrations. The resulting rotational atomic motions (chiral phonons) of predominantly in-plane F^- ions entails dynamic modulation of the crystal electric field. This results in polarization of the Ce^{3+} spins akin to the application of a strong magnetic field.



Department of Physics and Beus CXFEL Laboratory, Biodesign Institute, Arizona State University, Tempe, AZ, USA. Email: kaindl@asu.edu

dynamic control of magnetism with circularly polarized THz or mid-infrared pulses, indicating the generation of a magnetic moment in strontium titanate ($SrTiO_3$) (*12*) and evidencing magnetic switching of metallic nanolayers mediated by a dielectric substrate (*13*). Effective magnetic fields estimated in these studies ranged up to several tens of millitesla. In the work by Luo *et al.*, the chiral phonon–induced magnetization is found to correspond to a significantly larger effective magnetic field of ~0.9 T, as quantified from the magnetization dynamics.

Notably, the simple picture of the rotational motion of the ionic charges generating a magnetic field analogous to an electric loop current fails to account for experimental observations by several orders of magnitude. Consequently, additional physics is required to explain the large magnetic response of chiral phonons—both the magnetization arising from chiral phonons and the large splitting of corresponding phonon lines observed in external magnetic fields. For rare-earth trihalides, as applicable to CeF_3, theoretical models have linked the giant effective magnetic moments of chiral phonons to a mechanism whereby dynamic modulation of the crystal field by the rotating ionic displacements drives transitions between different orbital states of the central cation (*14*). The possibility of phonon-driven effective magnetic fields up to several tens of tesla has been suggested, motivating experimental exploration of these limits in strong-field conditions (*6, 15*).

Chiral phonons hence act as mediators of magnetization and angular momentum at the microscopic scale. Coupled to electromagnetic fields, they represent distinct control "knobs" that enable tailored energy flow—from light to phonons to spins—to ultimately steer materials into new nonequilibrium phases and spintronic functions. These exciting developments exploiting chiral phononics are thus set to write a new chapter in magnetism. ■

REFERENCES AND NOTES

1. A. Einstein, W. J. d. Haas, *Verh. Dtsch. Phys. Ges.* **17**, 152 (1915).
2. E. Beaurepaire *et al.*, *Phys. Rev. Lett.* **76**, 4250 (1996).
3. A. Kirilyuk, A. V. Kimel, T. Rasing, *Rev. Mod. Phys.* **82**, 2731 (2010).
4. C. Dornes *et al.*, *Nature* **565**, 209 (2019).
5. S. R. Tauchert *et al.*, *Nature* **602**, 73 (2022).
6. J. Luo *et al.*, *Science* **382**, 698 (2023).
7. H. Chen *et al.*, *2D Mater.* **6**, 012002 (2018).
8. H. Zhu *et al.*, *Science* **359**, 579 (2018).
9. H. Ueda *et al.*, *Nature* **618**, 946 (2023).
10. T. Zhang *et al.*, *Nano Lett.* **23**, 7561 (2023).
11. T. F. Nova *et al.*, *Nat. Phys.* **13**, 132 (2017).
12. M. Basini *et al.*, arXiv 2210.01690 (2022).
13. C. S. Davies *et al.*, arXiv 2305.11551 (2023).
14. S. Chaudhary *et al.*, arXiv 2306.11630 (2023).
15. D. M. Juraschek, T. Neuman, P. Narang, *Phys. Rev. Res.* **4**, 013129 (2022).

10.1126/science.adl3521

REPRODUCTION

A case of mistaken epigenetic identity

The specialized packaging of sperm DNA preserves genome stability in the fruit fly zygote

By **Mia T. Levine**

Paternal chromosomes undergo a radical conformational change during sperm development, compacting 20- to 200-fold (*1, 2*). Compaction is achieved by the replacement of canonical DNA packaging proteins with sperm-specialized versions. This specialized packaging renders paternal DNA biologically inert and unable to replicate or divide postfertilization in the zygote (*3*). To restore biological competency to the paternal DNA, canonical DNA packaging is reestablished. The biological importance of paternal chromosome compaction has been attributed largely to prefertilizaton sperm functions —for example, the need for superior hydrodynamic performance or the preservation of paternal genome integrity in the sperm head (*4, 5*). On page 725 of this issue, Dubruille *et al.* (*6*) report that specialized sperm DNA packaging also confers a vital, postfertilization function in fruit flies: the epigenetic identification of paternal chromosomes as distinct from maternal chromosomes. Without this epigenetic distinction, the paternal DNA shatters before the first zygotic division.

The epigenetic identity of sperm-deposited chromosomes is established during sperm maturation in the testis (*7, 8*). In the sperm of most animals, paternal chromosomes are almost devoid of nucleosomes, the spool-like structures around which DNA wraps. During spermiogenesis, sperm nuclear basic proteins (SNBPs) replace the nucleosome's octameric core, which is made up of a histone H2A-H2B tetramer and a histone H3-H4 tetramer. Once sperm DNA enters the egg, maternal proteins evict SNBPs and deposit histones (and histone variants). This de novo assembly of nucleosomes onto paternal DNA is vital for the replication and segregation of sperm-deposited DNA, and ultimately, transcription of the zygotic genome (*3*).

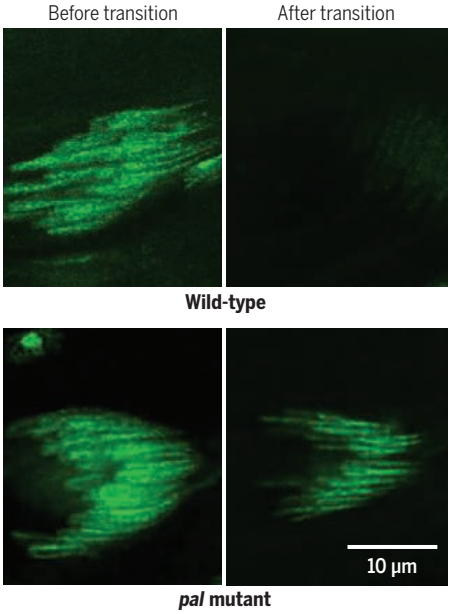
Despite the ubiquity of paternal chromosome compaction across animals, there is only a short list of characterized genes that regulate this developmental transition. These genes share a rare mutant phenotype: the development of fertilization-competent gametes but failure to produce live offspring. Most of these mutants have been identified by accident from genetic screens for regulators of sperm and egg development (*3, 9*).

One such serendipitous discovery emerged from a screen for autosomal genes that control meiosis in *Drosophila melanogaster*, conducted more than 50 years ago (*10*). A mutant line of flies termed paternal loss inducer (*pal*)—a name also given to the mutated gene—was identified. *pal* males fathered a reduced number of progeny. Among the progeny that did emerge ("escapers") were a subset that developed into adult flies but were lacking one or more paternal chromosomes. These "mosaic" escapers had patches of wild-type-looking cells next to patches of cells that appeared to lack a paternal chromosome (*11*). Typically, 50% of the cells had one genotype and 50% of the cells had another. It was inferred from this observation that paternal chromosome loss must occur not during meiosis but instead during the first embryonic division, when the animal generates its first two nuclei.

Dubruille *et al.* experimentally tested this prescient inference 45 years later. Using confocal microscopy to visualize the events immediately following fertilization, the authors detected paternal chromosome fragmentation in zygotes fathered by *pal* males. In these zygotes, the authors also observed that the maternal meiotic machinery, which normally segregates only the diploid maternal meiotic chromosomes, localized to the haploid paternal chromosomes as well. This aberrantly localized meiotic machinery shattered the paternal DNA. These data indicate that the paternal chromosomes are mistaken for maternal chromosomes in zygotes fathered by *pal* males, with catastrophic consequences.

To understand why the paternal chromosomes were ostensibly mistaken for maternal chromosomes, Dubruille *et al.* studied

Department of Biology, Epigenetics Institute, University of Pennsylvania, Philadelphia, PA, USA.
Email: m.levine@sas.upenn.edu



Drosophila melanogaster **spermatids are shown before and after the histone-to-protamine transition. Histone H3.2 (green) is removed from wild-type spermatids but not from *pal* mutant spermatids.**

sperm development. Rather than the normal wholesale eviction of nucleosomes from paternal DNA during sperm maturation, sperm from *pal* mutants retained the H3-H4 histone tetramer (see the figure), which, at fertilization, is normally found only on maternal chromosomes. The authors inferred that, after fertilization, the aberrantly retained paternal H3-H4 recruits maternal meiotic machinery to haploid paternal chromosomes.

The data collected from both sperm and embryos motivated a compelling prediction: The paternal DNA deposited by *pal* sperm, which aberrantly retains histones, should bypass the need for de novo histone deposition in the zygote. To test this, the authors bred *pal* males with females carrying a homozygous loss of the gene *Hira* (*12*). In the zygote, maternal HIRA deposits an H3 variant into the paternal DNA. When bred with wild-type males, mothers lacking HIRA give rise to embryos whose paternal chromosomes fail to reestablish the canonical packaging required for integration into the embryo. All of the embryos die. Remarkably, breeding mothers lacking HIRA with fathers lacking PAL yielded live offspring, indicating that sperm from *pal* mutants bypass the requirement for de novo histone deposition by HIRA. This elegant experiment confirms that *pal* sperm aberrantly retain histone H3-H4.

The discovery of a postfertilization function for paternal chromosome compaction provides a foundation to further probe sperm chromatin biology. Dubruille *et al.* found that PAL localizes to the sperm nucleus during the replacement of histones with SNBPs; however, how PAL promotes histone H3-H4 eviction during sperm maturation is unclear. Does PAL interact directly with histone H3-H4 or with another protein that interfaces directly with histones? *pal* is not found in mammals, so it will be important to identify the mammalian equivalent. These data also suggest that other maternal proteins in the embryo might exploit maternal-paternal epigenetic asymmetry to promote paternal DNA competency.

Paternal DNA integration into the embryo is essential for viability, yet Dubruille *et al.* report rapid evolution of *pal* across closely related *Drosophila* species. *pal* evolves faster than all other characterized sperm chromatin components in *Drosophila*. This discovery raises the possibility of recurrently changing histone abundance in mature sperm over short evolutionary timescales. Sequence evolution of only a single gene like *pal* may be sufficient to alter sperm histone retention, with consequences for sperm competition and zygotic gene expression. A deeper investigation of histone retention differences across closely related species, combined with molecular evolution analysis of genes such as *pal* that regulate histone retention, would offer insights into the biological pressures and mechanisms that shape the evolution of sperm chromatin composition. Notably, the findings of Dubruille *et al.* suggest that this evolution is bounded. Tuning histone retention in response to prefertilization pressures such as sperm competition may trade off against postfertilization pressure to ensure a distinct, paternal chromosome epigenetic identity. ■

REFERENCES AND NOTES

1. R. E. Braun, *Nat. Genet.* **28**, 10 (2001).
2. M. Fuller, in *The Development of Drosophila melanogaster*, M. B. a. A. Martinez-Arias, Ed. (Cold Spring Harbor Laboratory Press, 1993), pp. 71–147.
3. B. Loppin, R. Dubruille, B. Horard, *Open Biol.* **5**, 150076 (2015).
4. C. Cho *et al.*, *Biol. Reprod.* **69**, 211 (2003).
5. L. Lüke, P. Campbell, M. Varea Sánchez, M. W. Nachman, E. R. S. Roldan, *Proc. Biol. Sci.* **281**, 20133359 (2014).
6. R. Dubruille *et al.*, *Science* **382**, 725 (2023).
7. L. Moritz, S. S. Hammoud, *Front. Endocrinol.* (Lausanne) **13**, 895502 (2022).
8. C. Rathke, W. M. Baarends, S. Awe, R. Renkawitz-Pohl, *Biochim. Biophys. Acta. Gene Regul. Mech.* **1839**, 155 (2014).
9. K. R. Fitch, G. K. Yasuda, K. N. Owens, B. T. Wakimoto, *Curr. Top. Dev. Biol.* **38**, 1 (1998).
10. L. Sandler, *Drosoph. Inf. Serv.* **47**, 68 (1971).
11. B. S. Baker, *Genetics* **80**, 267 (1975).
12. B. Loppin *et al.*, *Nature* **437**, 1386 (2005).

ACKNOWLEDGMENTS

The author thanks A. Das for critically reading an earlier draft. M.T.L. was supported by National Institutes of Health grant R35GM124684.

10.1126/science.adl0365

MATERIALS SCIENCE

Staying stably cool in the sunlight

Microporous ceramics passively cool buildings and reduce the need for air conditioners

By **Dongliang Zhao[1,2]** and **Huajie Tang[1]**

Terrestrial surfaces exposed to sunlight absorb solar heat and shed heat back to outer space as infrared radiation. If the radiated heat is greater than the solar energy absorbed, then daytime radiative cooling is achieved passively, without any energy input. However, this approach requires materials that strongly reflect sunlight and simultaneously emit long-wavelength infrared light—the wavelength needed to escape Earth's atmosphere and not be reflected back. Ceramics composed mainly of silica (SiO_2) and alumina (Al_2O_3), and that permit long-wavelength infrared emission, can meet these requirements. Moreover, they are durable, thermostable, and water-resistant. On page 684 and 691 of this issue, Zhao *et al.* (*1*) and Lin *et al.* (*2*), respectively, describe microporous materials—a glass-based ceramic coating and a ceramic composite—that exhibit passive daytime radiative cooling and resistance to harsh environments. These advances may lead to more environmentally friendly ways for keeping buildings cool.

Nanophotonic cooler design opened the way for daytime radiative cooling, but engineering nanostructured materials to reflect and emit light at specific wavelengths is costly (*3*). To achieve low-cost and scalable manufacturing, polymer-based structures were fabricated, including a polymer-metal hybrid film (*4*), porous polymer (*5*), and polymer-dielectric paint (*6*). However, organic polymers are prone to aging and exhibit poor durability outdoors. Ceramics such as SiO_2, silicon carbide (SiC), and

[1]School of Energy and Environment, Engineering Research Center of Building Equipment, Energy, and Environment, Southeast University, Nanjing 210096, China. [2]Institute of Science and Technology for Carbon Neutrality, Southeast University, Nanjing 210096, China.
Email: dongliang_zhao@seu.edu.cn

Al_2O_3 have outstanding environmental stability. Although ceramic-based fiber membranes have been demonstrated for radiative cooling (*7*), it has been challenging to fabricate ceramic structures with sufficient mechanical strength.

The light transmission properties of a material are dictated by its refractive index and extinction coefficient. The refractive index indicates how much the light's trajectory deviates when crossing a material; the extinction coefficient is proportional to the amount of light absorbed by the material. Ideal materials for daytime radiative cooling generally require a high refractive index, a low extinction coefficient in the solar spectrum (near-infrared, visible, and ultraviolet), and a high extinction coefficient for the long-wavelength infrared light. However, materials with both a low extinction coefficient in the solar spectrum and a high extinction coefficient in the long wavelength infrared, such as polymers and metal oxides, usually possess low refractive indexes in the solar spectrum.

An effective solution to this conundrum requires tailoring the solar reflectivity and absorptivity of a material by resonantly scattering the incident light. Accordingly, silica and alumina particles can be ideal because they resonantly scatter light in the range of wavelengths close to their diameters with the highest efficiency (Mie scattering). Ceramics have extremely low extinction coefficients in the solar spectrum. The latter characteristic is determined by the electronic structure of the ceramic molecule. When the energy of incident light is insufficient to excite electrons in silica or alumina molecules (because the electronic bandgap is too great), the material does not exhibit absorbance. Thus, resonant scattering by alumina and silica just enhances solar reflectivity, not absorptivity. Outside of the solar spectrum, however, these ceramics possess a high extinction coefficient, so the resonance effect enhances the emission of long-wavelength infrared radiation. By optimizing the microstructure of ceramics, Zhao *et al.* and Lin *et al.* demonstrated a near-perfect solar reflectivity of more than 96 and 99.6% and a high emissivity (of long-wavelength infrared radiation) of more than 95 and 96.5%, respectively.

The cooling glass–based ceramic coating developed by Zhao *et al.* comprises a microporous SiO_2-based framework embedded with Al_2O_3 nanoparticles, whereas the cooling ceramic composite developed by Lin *et al.* is composed of a microporous Al_2O_3 framework. Both involve solar light scattering on the structure of Al_2O_3. However, Zhao *et al.* optimized Al_2O_3 particles as the resonators, whereas Lin *et al.* achieved light scattering from within the pores of Al_2O_3 framework. Zhao *et al.* also modulated the diameters of the SiO_2 constituents of the glass particles to match long-wavelength infrared radiation (which thus facilitates long-wavelength infrared emission). Although the two cooling materials differ in structure, their underlying design principles are similar (see the figure).

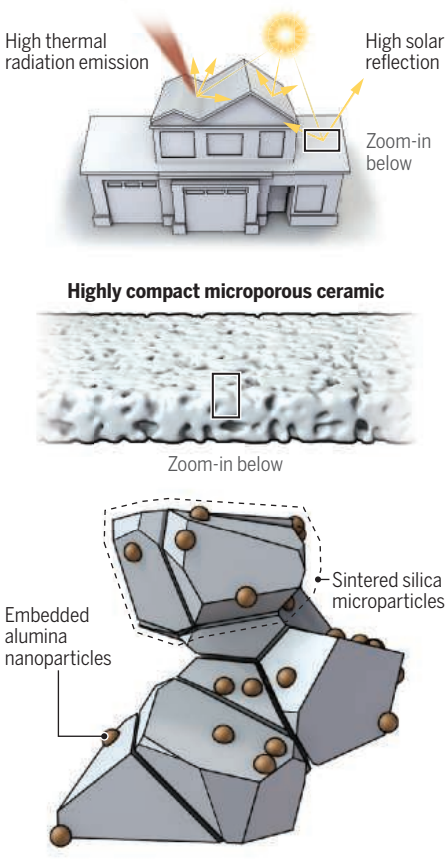
In addition to their absorption characteristics, the respective trigonal SiO_2 and hexagonal Al_2O_3 close-packed crystal structures guarantee their environmental stability and resistance to heat, pressure, acid, and alkali conditions. Zhao *et al.* conducted several durability tests, including exposures to flame shock, water immersion, and ultraviolet radiation for the cooling glass–based ceramic coating. Lin *et al.* performed a test simulating the calcination and liquid droplet evaporation of the cooling ceramic composite under fire conditions. Both studies also report that solar reflectivity did not change after the materials were exposed to outdoor environments, and both materials resisted soiling under hot and humid climates. Similar to the glazing process applied to porcelain surfaces, Zhao *et al.* applied a thin and dense transparent glass layer onto the cooling glass–based ceramic coating to enhance its resistance to liquid pollutants such as rainwater mixed with dust and to prevent airborne dust ingress into the microporous silica framework.

Both Zhao *et al.* and Lin *et al.* prepared precursors for their cooling materials by dispersing ceramic particles in organic solvents and then sintering the mixtures to remove organic impurities and moisture. This method achieved the desired structures and mechanical strengths. However, for real-world applications, conventional sintering must be integrated with existing material production processes for construction materials. For example, the cooling glass–based ceramic coating can be sintered on roofing tiles by using "roll-to-roll" industrial processes in which materials are continuously produced on a roll from a flexible substrate. When roofing a house with the cooling ceramic composite, Lin *et al.* determined that 26.8% of the electricity used for air conditioning was saved in hot and humid cities such as Hong Kong. Zhao *et al.* estimated that the annual CO_2 emissions associated with air conditioning use could be reduced by 2.1 tons (~8.3%) and 1.6 tons (~7.6%) for old midrise apartments (pre-1980) and new midrise apartments (post-2004), respectively, in US cities.

Strategies to reduce carbon emissions through the use of cooling ceramics should consider the durability and life cycle of these materials rather than focusing mainly on energy savings and cost effectiveness during their use (*8*). Although some structures with dynamic radiative cooling capabilities have been proposed and experimentally demonstrated recently (*9*), attaining large-scale applications remains a substantial challenge. Nevertheless, the findings of Zhao *et al.* and Lin *et al.* advance cooling approaches that could, if commercially applied to buildings, drive down the electrical demand of air conditioners and benefit the environment. ■

A cool building design

Materials, including roofing tiles, could be coated with a microporous ceramic (such as silica and alumina) that strongly reflects sunlight and emits thermal radiation (in the long-wavelength infrared spectrum). For example, sintering silica microparticles creates a porous material (which also accommodates embedded alumina nanoparticles) that can stably and passively cool buildings during the day.



GRAPHIC: A. FISHER/*SCIENCE*

REFERENCES AND NOTES

1. X. Zhao *et al.*, *Science* **382**, 684 (2023).
2. K. Lin *et al.*, *Science* **382**, 691 (2023).
3. A. P. Raman *et al.*, *Nature* **515**, 540 (2014).
4. Y. Zhai *et al.*, *Science* **355**, 1062 (2017).
5. J. Mandal *et al.*, *Science* **362**, 315 (2018).
6. J. Mandal *et al.*, *Joule* **4**, 1350 (2020).
7. T. Li *et al.*, *Chem. Eng. J.* **452**, 139518 (2023).
8. H. Tang *et al.*, *J. Clean. Prod.* **380**, 135035 (2022).
9. S. Wang *et al.*, *Science* **374**, 1501 (2021).

10.1126/science.adk9614

HYPOTHESIS

How do molecular motors fold the genome?

A potential mechanism of DNA loop extrusion by molecular motors is discussed

By **Cees Dekker,[1] Christian H. Haering,[2] Jan-Michael Peters,[3] Benjamin D. Rowland[4]**

Many protein complexes that drive key processes in cells are "molecular motors"—assemblies that consume (electro)chemical energy to produce mechanical work. Examples include the F_oF_1 synthase rotary motor that catalyzes adenosine triphosphate (ATP) production, kinesin and myosin motors that "walk" along cytoskeleton filaments, or polymerases and helicases that move along DNA. Structural-maintenance of chromosomes protein complexes (SMCs) have only recently been identified as an entirely distinct class of DNA-translocating motors, although their key role in folding the linear DNA double helix into intricate three-dimensional structures, such as X-shaped mitotic chromosomes, was known for decades. Here, we discuss how insights from biophysical, biochemical, and structural studies are starting to yield an understanding of the mechanism by which these motors extrude loops of DNA to structure genomes.

SMCs are evolutionarily conserved from bacteria to humans. Eukaryotes feature three main classes of SMCs: condensin, which assembles mitotic chromosomes during cell division; cohesin, which regulates interphase chromosome structure and links sister chromatids (the two copies of every chromosome generated by DNA replication); and SMC5/6, which has less well understood roles in DNA damage repair and replication. All of these complexes exhibit a similar tripartite ring architecture of ~40 nm in diameter made of a dimer of coiled-coil SMC proteins and an intrinsically disordered kleisin protein, to which additional subunits attach. In the case of cohesin and condensin, these additional subunits are built from multiple repeats of "HEAT" motifs, referred to here as HEAT-A and HEAT-B subunits. At the heart of the motor are two globular ATPase head domains located at the ends of the SMC coiled coils.

SMCs generate DNA loops, which appear to be the basic motif of chromosome structure that underlie many major genomic processes, from folding mitotic chromosomes to the formation of topologically associating domains [TADs (genomic domains thought to regulate gene expression)]. Whereas earlier chromosome conformation capture mapping and polymer simulations suggested that extrusion of DNA loops by SMCs could explain many chromosomal features, direct evidence for such loop extrusion by SMCs was provided by single-molecule studies that visualized the formation of DNA loops in real time (*1*–*4*).

These studies yielded a wealth of data. Driven by ATP hydrolysis, SMCs were found to be very fast motors, reeling in DNA at a speed of ~1 kilobase pair per second in a directional and processive manner for long distances. ATP binding induces SMCs to take a step of hundreds of base pairs—strikingly different from previously characterized DNA-translocating motors that typically move a single base pair at a time. Such large steps are consistent with studies that implicated conformational changes of SMC structure that were the approximate size of the entire complex in the DNA loop extrusion process (*5*–*8*). Although fast, SMCs are also weak motors that stall if subpiconewton forces are applied to the DNA that they reel in. Another unexpected feature of SMCs is their ability to pass DNA binding proteins, such as nucleosomes, polymerases, or even other SMCs, and incorporate them into the extruded DNA loops (*9*, *10*).

The data from these single-molecule studies sparked many debates about the mechanism of DNA loop extrusion by SMCs (*11*). Based on the resemblance of the coiled-coil architecture of SMC proteins to kinesin or myosin motors that transport cargo along cytoskeletal filaments, it was initially suggested that the two globular ATPase heads might grab two distal DNA sections and bring them together, thus creating a DNA loop. This intuitive early scenario appeared, however, incompatible with the small head-to-head distances consistently observed in structural studies of SMCs.

Alternative "scrunching" models suggested that the base of a DNA loop is clamped in the vicinity of the ATPase heads, but the hinge domain at the other end of the coiled coils reaches out to grab a new DNA segment. Bending of the coils, driven by thermal fluctuations, "swings" the hinge with the new DNA section into the vicinity of the heads, where it is handed over from the hinge to the heads to merge with the DNA loop (*6*). The reverse order is also conceivable, with a DNA section bound at the hinge domain moving away from the ATPase heads while the coiled coils straighten (*12*).

Yet another hypothesis is based on the idea that opening up the lumen of the SMC-kleisin ring allows the capture of a DNA loop segment, which could be generated either by thermal fluctuations (*13*) or actively fed into the lumen by a "power-stroke" motion that is driven by the energy of ATP binding to the ATPase heads (*14*, *15*). Similar to the scrunching-type models, the newly captured loop then merges with a preexisting DNA loop held in the vicinity of the ATPase heads when the coiled coils bend and reassociate.

The different models have been tested against various experimental data: Coil bending was observed by atomic force microscopy (*5*, *6*) and cryo–electron microscopy (cryo-EM) (*7*, *8*) for multiple SMCs, predominantly at a local discontinuity in the coiled-coil superhelix that acts as an "elbow." Locking the two SMC subunits in the bent conformation prevented loop extrusion (*6*). In cryo-EM structures, DNA was bound to the inner surfaces of the SMC-kleisin ring lumen (*11*), arguing that the ring might topologically encircle DNA strands (in a manner that either protein or DNA would have to be cleaved to disengage the two) or pseudo-topologically embrace a DNA loop (in which case DNA could be pulled out of the protein ring without severing either of the two) (*15*).

Interestingly, fusing all SMC-kleisin interfaces to establish one continuous cohesin ring did not prevent loop extrusion (*2*, *9*), which argued against a classical topological model in which the SMC-kleisin ring would have to be opened to entrap DNA. Furthermore, the finding that obstacles larger than the cohesin ring could be incorporated into DNA loops (*9*) supported a nontopological model in which the extruded loop is not embraced by the SMC-kleisin lumen. The seemingly contradictory conclusions on the topology could be reconciled upon redrawing parts of the kleisin's path relative to the DNA (*11*), which implied that DNA can be temporarily cap-

[1]Department of Bionanoscience, Kavli Institute of Nanoscience Delft, Delft University of Technology, Delft, Netherlands. [2]Department of Biochemistry and Cell Biology, Julius Maximilian University of Würzburg, Würzburg, Germany. [3]Research Institute of Molecular Pathology (IMP), Campus-Vienna-Biocenter 1, Vienna, Austria. [4]Division of Cell Biology, The Netherlands Cancer Institute, Amsterdam, Netherlands. Email: c.dekker@tudelft.nl; christian.haering@uni-wuerzburg.de; peters@imp.ac.at; b.rowland@nki.nl

tured pseudo-topologically within the SMC-kleisin ring during the ATP hydrolysis cycle, whereas the final extruded loop is not topologically embraced by the SMC-kleisin ring.

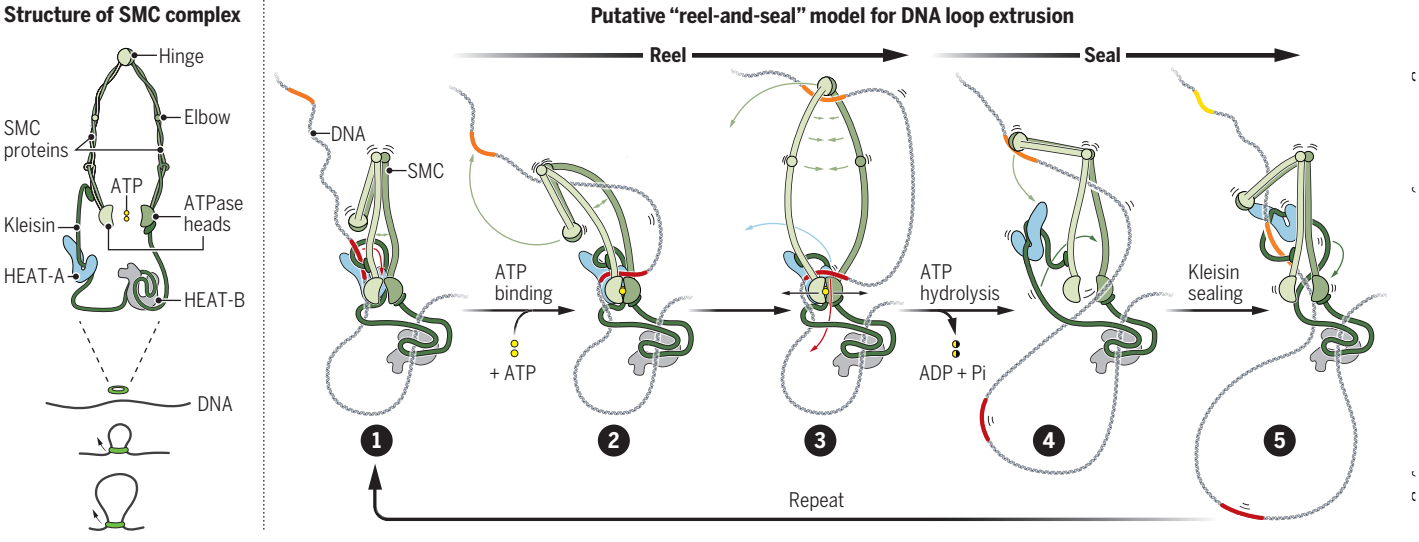
Although there are further differences between the various types of models—for example, in the way that force is generated— they all rely on large conformational rearrangements to relocate a stretch of DNA in a single reaction cycle. Below, we attempt to put together a hybrid working model that integrates central features of the different hypotheses. We base our speculations on the head engagement forms a new DNA binding site on the SMC heads and initiates a series of large-scale movements that involve a power stroke: Presumably assisted by a conformational change in HEAT-A, DNA gets clamped onto the SMC heads, which triggers a stretching and opening up of the SMC arms. These movements swing the SMC hinge away from the ATPase heads, de facto capturing a looped DNA segment inside the lumen between the disengaged arms. Clamping of the DNA onto the heads and enlarging the loop by reeling a DNA segment through the SMC lumen pro-

the SMC-kleisin lumen. Reel-and-seal extends beyond a series of previous models (*11*, *13*, *15*) by invoking the motion that unseals the HEAT-A kleisin fold, which provides the window of opportunity for the complex to pass obstacles that can be substantially larger than the diameter of the ring (*9*). Although this hypothetical scenario is compatible with most currently available data, it is intended to represent an intellectual exercise to inspire much-needed further studies.

Several key questions need to be addressed to fully elucidate the mechanism through

DNA loop extrusion by structural maintenance of chromosomes complexes

SMCs such as cohesin and condensin have a ring architecture that consists of two SMC coiled-coil subunits and a disordered kleisin subunit that is bound by HEAT-A (in humans: NIPBL for cohesin, NCAPD2 or NCAPD3 for condensin) and by HEAT-B (in humans: STAG1 or STAG2 for cohesin, NCAPG or NCAPG2 for condensin) subunits. In our proposed "reel-and-seal" model for DNA loop extrusion by SMCs, the interfaces of the SMC-kleisin ring do not need to disengage for the complex to accommodate DNA. Instead, the protein complex folds around the DNA (**1**). Upon ATP binding to the folded resting state, DNA gets clamped onto the engaged ATPase heads (**2**), which induces stretching and opening of the coiled coils and transfer of a new DNA loop between the disengaged coils (**3**). Upon ATP hydrolysis, the newly formed DNA loop is transferred into the extruded DNA loop (**4**) and the SMC relaxes into its folded state (**5**). Red, orange, and yellow indicate DNA segments that are sequentially transferred into the extruded loop.



ADP, adenosine diphosphate; ATP, adenosine triphosphate; NCAPD2, non-SMC condensin subunit D2; NCAPD3, non-SMC condensin subunit D3; NCAPG, non-SMC condensin subunit G; NCAPG2, non-SMC condensin subunit G2; NIPBL, nipped-B-like protein; Pi, inorganic phosphate; SMC, structural maintenance of chromosomes complex; STAG1, cohesin subunit SA-1; STAG2, cohesin subunit SA-2.

two best-studied SMCs: cohesin and condensin. As key features of SMC action appear to be universal across all the complexes, it will be interesting to see whether SMC5/6 or prokaryotic SMCs use a similar mechanism.

In our hybrid model (see the figure and online movie), loop-extruding SMCs do not topologically embrace DNA inside their ring-shaped structure, but rather fold their peptide chain around the DNA. The two HEAT-repeat subunits each wrap part of the long kleisin subunit around the DNA. The fold at HEAT-B provides an anchor, which allows the complex to hold onto a DNA section while enlarging the DNA loop in one direction at any given time. The second fold, at HEAT-A, positions the DNA double helix such that it can be fed onto the SMC ATPase heads upon ATP binding. ATP binding–induced SMC vides directionality to the translocation. At this point, the hinge domain contacts a new DNA section. In the next step, ATP hydrolysis and nucleotide release trigger HEAT-A to leave the ATPase heads, and the arms bend over again, which causes the segment of the captured DNA loop between the arms to move toward the heads. In this step, the original DNA section is transferred through the ATPase heads to end up in the extruded DNA loop that consequently further grows in size. The new DNA section is sealed into the HEAT-A kleisin fold and is thus positioned for clamping on top of the ATPase heads in the next cycle.

This "reel-and-seal" model explains how DNA transiently inserts a pseudotopological loop into the SMC-kleisin ring, while the final extruded DNA loop is not embraced within which SMCs extrude DNA into loops. How DNA is bound by and translocated through SMCs by a combination of topological and direct binding interactions needs to be clarified, as does the question whether translocation requires cycles of opening and closing of the SMC arms and/or cycles of stretching and bending of the arms. Whether the DNA "slides" or is handed-over from one binding site to another is also not yet clear. In addition, it will be important to understand how SMC conformational changes relate to functions other than loop extrusion, such as sister chromatid cohesion or translocation of SMCs along DNA.

Other questions concern the functions of the HEAT-repeat subunits. Whether HEAT-A might contact the SMC hinge of condensin as it does for cohesin, and

GRAPHIC: DEKKER *ET AL.*

whether HEAT-B might have a DNA-anchoring role for cohesin as it does for condensin are still unknown. An exchange between cohesin's HEAT-A variants nipped-B-like protein (NIPBL) and PDS5 may control DNA looping, but this mechanism is yet to be understood. Another outstanding question is whether SMCs extrude DNA loops unidirectionally or whether they can switch directionality by exchanging the bound DNA segments between HEAT-A and HEAT-B.

Because the looping of DNA by SMCs might change the twist of the DNA double helix, it will be important to understand the role of this induced torsion, which will supercoil the extruded DNA loop. Further insight into how the forces are generated that translocate and bend the DNA during loop extrusion will also be required. It is also not yet clear whether the energy released during ATP binding-hydrolysis cycles is used to generate these forces, or whether SMCs function as molecular ratchets that are driven by thermal motion.

Similarly, it will be of great importance to understand how SMCs can operate on chromatin fibers, where DNA is occupied by nucleosomes and polymerases. Whether SMCs step from one accessible DNA region to the next, or whether they can interact with DNA while it is wrapped around histone octamers, needs to be addressed. How cohesin's loop extrusion cycle is controlled by the architectural protein CTCF that defines the boundaries of TADs, and how post-translational modifications such as acetylation or phosphorylation of different SMC subunits affect loop extrusion, are other questions of interest. It will also be important to understand whether SMC5/6 and bacterial SMCs act differently than condensin and cohesin and whether protein complexes of similar architecture, such as the MRE11-RAD50-nibrin DNA repair protein complex, also extrude DNA. ■

REFERENCES AND NOTES

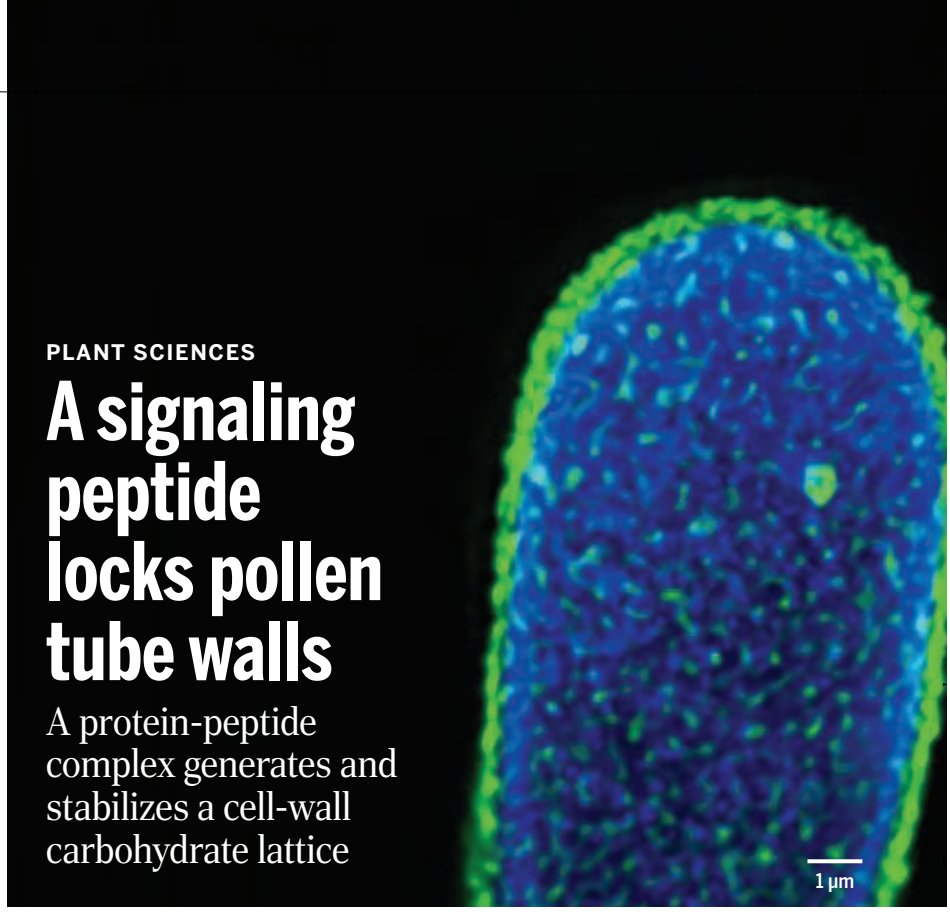
1. M. Ganji *et al.*, *Science* **360**, 102 (2018).
2. I. F. Davidson *et al.*, *Science* **366**, 1338 (2019).
3. Y. Kim, Z. Shi, H. Zhang, I. J. Finkelstein, H. Yu, *Science* **366**, 1345 (2019).
4. B. Pradhan *et al.*, *Nature* **616**, 843 (2023).
5. J.-K. Ryu *et al.*, *Nat. Struct. Mol. Biol.* **27**, 1134 (2020).
6. B. W. Bauer *et al.*, *Cell* **184**, 5448 (2021).
7. B.-G. Lee *et al.*, *Nat. Struct. Mol. Biol.* **27**, 743 (2020).
8. F. Bürmann *et al.*, *Nat. Struct. Mol. Biol.* **26**, 227 (2019).
9. B. Pradhan *et al.*, *Cell Rep.* **41**, 111491 (2022).
10. E. Kim, J. Kerssemakers, I. A. Shaltiel, C. H. Haering, C. Dekker, *Nature* **579**, 438 (2020).
11. R. Oldenkamp, B. D. Rowland, *Mol. Cell* **82**, 1616 (2022).
12. T. L. Higashi, G. Pobegalov, M. Tang, M. I. Molodtsov, F. Uhlmann, *eLife* **10**, e67530 (2021).
13. J. F. Marko, P. De Los Rios, A. Barducci, S. Gruber, *Nucleic Acids Res.* **47**, 6956 (2019).
14. S. K. Nomidis, E. Carlon, S. Gruber, J. F. Marko, *Nucleic Acids Res.* **50**, 4974 (2022).
15. I. A. Shaltiel *et al.*, *Science* **376**, 1087 (2022).

10.1126/science.adi8308

PLANT SCIENCES

A signaling peptide locks pollen tube walls

A protein-peptide complex generates and stabilizes a cell-wall carbohydrate lattice



Demethylesterified pectin (green) forms a reticulated pattern in wild-type *Arabidopsis thaliana* pollen tubes.

By **Debra Mohnen**

Pollen tube elongation in flowering plants transports sperm through the stigma and style to the egg in the ovary, enabling fertilization. Pollen tube growth entails oscillatory turgor-driven polarized cell wall expansion, which requires a malleable cell wall enriched in specific carbohydrates and proteins (*1*). However, what all of these carbohydrates and proteins are, and how they enable growing pollen tubes to elongate at rates of up to 1 cm hour^{-1} (*2*) while providing sufficient strength to prevent cell rupture, is not clear. Growing pollen tubes have a pectin-containing lattice in the wall behind the growing pollen tube tip (*3*). Many of the components of this lattice and how its formation is regulated are unknown. On page 719 in this issue, Moussu *et al.* (*4*) identify an unexpected player in this process: a peptide that forms a complex with a structural protein and binds deesterified pectin to generate and stabilize a reticulate tripartite network architecture in the pollen tube wall.

Pectin is a complex glycan found in the plant cell wall (*5*). Pollen tube growth occurs in the first 20 µm of the growing tip in a clear zone devoid of starch grains and vacuoles and enriched in vesicles containing methylesterified pectic homogalacturonan (HG) that provides a malleable wall for tip growth (*1*). HG, the most abundant pectic glycan in growing cells (*5*), is a galacturonic acid–rich homopolymer that is highly negatively charged if it is not methylesterified. The tubular lattice present behind the growing tip is enriched in Ca^{2+}–cross-linked deesterified HG that, along with cellulose and callose, strengthens the pollen tube wall (*2*, *6*). Cell wall–localized pectin methylesterases (PMEs) deesterify newly synthesized HG after delivery to the pollen tube tip (*1*), generating carboxylate groups in the HG chain that are hypothesized to sequester Ca^{2+} from preexisting HG, enabling simultaneous tip growth, wall strengthening, and avoidance of cell bursting (*7*). Transmembrane flux of additional Ca^{2+} into the wall further strengthens the HG-Ca^{2+}-HG network behind the new area of expansion.

Moussu *et al.* used elongating pollen tubes from *Arabidopsis thaliana* to investigate how cell wall polymers such as HG generate the architectural patterns that support plant cell growth. They focused on RAPID ALKANIZATION FACTOR 4 (RALF4), a member of the RALF signaling peptide family, which regulates plant growth, development, and response to

University of Georgia, Athens, GA, USA.
Email: dmohnen@ccrc.uga.edu

pathogens by interacting with receptor kinases (*8*, *9*). It was previously shown that a mutation in the gene encoding RALF4 leads to misshapen pollen tubes, tube bursting, and cell wall composition changes (*10*). In addition, RALF4 forms a tetrameric complex with cell wall LEUCINE-RICH REPEAT EXTENSIN 8 (LRX8), which exposes a positively charged patch on RALF4 that is hypothesized to interact with negatively charged pectin (*11*). RALF4 also binds to members of the *Catharanthus roseus* RLK1-like family of receptor kinases and influences pollen tube function (*10*, *12*). On the basis of these multiple activities, Moussu *et al.* hypothesized that RALF4 triggered two parallel signaling pathways in the pollen tube: one inside the cell through receptor kinases and one outside the cell through interaction with leucine-rich extensins (*11*).

Using thermal shift assays, Moussu *et al.* showed that two RALF4 peptides form a complex with two LRX8 proteins and that the complex binds two deesterified HG molecules, one for each RALF4 subunit. Immunolabeling and high-resolution microscopy revealed that LRX8 and RALF4 are present both in the pollen tube tip, where they are exocytosed in vesicles into the tip, and also in the pollen tube shank, demonstrating that both are incorporated into the shank wall. This was unexpected, given that RALF4 is a signaling molecule. Super-resolution microscopy identified the heterotetrameric LRX8-RALF4 as punctate domains in a reticulated network in the wall. Deesterified pectin was present in the pollen tube wall as interconnected fibers (see the image) that colocalized with RALF4 in the reticulate network, reminiscent of the latticed and ring structures previously identified in pollen tubes (*3*).

To determine whether the association between LRX8-RALF4 and the pectin-labeled fibers occurred on preexisting fibers or as an active process in which LRX8-RALF4 influenced pectin fiber architecture, Moussu *et al.* analyzed the interaction of the LRX8-RALF4 complex with pectin. They showed that the ionic binding of RALF4 in the LRX8-RALF4 complex with pectin condensed demethylesterified pectin. Furthermore, compared with wild-type pollen tubes, pollen tubes with RALF4 that were mutated to remove the positively charged patch (to prevent the formation of the LRX8-RALF4 complex) displayed a relatively unstructured cell wall, grew at a greater velocity, and had premature tube bursting. This shows that the interaction of the LRX8-RALF4 complex with deesterified pectin is necessary for pollen tube patterning, growth, and integrity.

The study by Moussu *et al.* provides evidence that a protein-peptide-pectin interaction establishes and stabilizes a reticulate cell wall architecture in pollen tube walls that strengthens the pollen tube during growth. The results establish RALF4 as a multifunctional peptide with both signaling and structural roles in pollen tubes. A critical question for the future is to determine whether RALF4 or other RALF peptides generate cell wall architecture in non–tip growing cells, such as elongating epidermal or seedling stem hypocotyl cells. It will also be important to establish whether other RALF peptides have a dual signaling and structural role.

Another outstanding question regards the full polymeric structure of pectin in the lattice. There is little evidence that HG is an independent polysaccharide in plant cell walls. Rather, HG is covalently connected to the pectin rhamnogalacturonan I (RG-I) or to a pectic arabinogalactan protein proteoglycan and/or contains regions of the highly branched pectic RG-II (*5*, *13*, *14*). To understand how the LRX8-RALF4 stabilized lattice affects pollen tube shape, strength, and elongation, the structure of the pectic fibers in the lattice must be determined. Moussu *et al.* have identified RALF4 as a peptide that has a structural role in the cell wall and a signaling role in response to cell wall integrity fluctuations. This establishes RALF4 as a sentinel for rapid communication between the cell and the cell wall when rapid changes in cell wall architecture are required. ■

REFERENCES AND NOTES

1. P. K. Hepler, C. M. Rounds, L. J. Winship, *Mol. Plant* **6**, 998 (2013).
2. P. A. Bedinger, K. J. Hardeman, C. A. Loukides, *Trends Cell Biol.* **4**, 132 (1994).
3. J. Derksen *et al.*, *Plant J.* **68**, 495 (2011).
4. S. Moussu *et al.*, *Science* **382**, 719 (2023).
5. M. A. Atmodjo, Z. Hao, D. Mohnen, *Annu. Rev. Plant Biol.* **64**, 747 (2013).
6. C. Cameron, A. Geitmann, *Curr. Opin. Genet. Dev.* **51**, 11 (2018).
7. T. E. Proseus, J. S. Boyer, *Plant Signal. Behav.* **7**, 1490 (2012).
8. G. Pearce, D. S. Moura, J. Stratmann, C. A. Ryan Jr., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12843 (2001).
9. M. Haruta, G. Sabat, K. Stecker, B. B. Minkoff, M. R. Sussman, *Science* **343**, 408 (2014).
10. M. A. Mecchia *et al.*, *Science* **358**, 1600 (2017).
11. S. Moussu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7494 (2020).
12. Z. Ge *et al.*, *Science* **358**, 1596 (2017).
13. L. Tan *et al.*, *Carbohydr. Polym.* **301**, 120340 (2023).
14. F. Dardelle *et al.*, *Plant Physiol.* **153**, 1563 (2010).

ACKNOWLEDGMENTS

The work of D.M. is partially funded by The Center for Bioenergy Innovation, a US Department of Energy (DOE) Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science and the DOE, Office of Science, Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences Division, under award DE-SC0015662.

10.1126/science.adl1198

INFECTIOUS DISEASE

Preparing for the next pandemic

New lead drugs to treat COVID-19 are beginning to emerge

By **Brian K. Shoichet** *and* **Charles S. Craik**

Early in the COVID-19 pandemic, it was suggested that drugs to meaningfully treat the disease would come from either the current antiviral armamentarium or from drug repurposing (*1*). Genuinely new therapeutics, although likely to be the most effective, would take too long to develop. As it happened, drug repurposing paid few lasting dividends (*2*, *3*), and the effective drugs that did emerge, such as remdesivir and nirmatrelvir [the active ingredient in Paxlovid that targets the main protease (Mpro) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)], were drawn from preexisting antiviral pipelines. On page 663 of this issue, Boby *et al.* (*4*) report an open science effort to discover new Mpro drugs to treat COVID-19. Such work could help develop treatments for emerging drug-resistant variants of SARS-CoV-2 (*5*) and other coronaviruses that are likely to emerge in the future.

The effort described by Boby *et al.* began early in the pandemic and combined innovations in the techniques and organization of drug discovery. The scientific innovations included high-throughput crystallographic fragment screening: The structures of more than 70 fragments—which are about a third to a half of the size of a druglike inhibitor—bound to the SARS-CoV-2 enzyme Mpro were determined (*6*) as were the structures of more than 500 more-advanced compounds. A second area of innovation was computational free-energy calculations and machine learning that helped guide medicinal chemistry; both techniques remain at the cutting edge of drug development.

Organizationally, the study of Boby *et al.* resulted from the integration of the expertise of 212 scientists in 47 organizations across 15 countries, mostly in academia but also including researchers from pharmaceuti-

Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA.
Email: charles.craik@ucsf.edu

cal companies. All donated their time in an open access model, capturing some of the strengths of a vertically integrated pharmaceutical company in a collaborative academic environment. They report several lead compounds with cellular antiviral potencies that are comparable to those of nirmatrelvir with in vivo pharmacokinetics that may support advancement toward clinical trials.

By focusing on Mpro, the authors drew on a legacy of successful drug discovery against viral proteases. Beginning in the 1980s and 1990s, investigators began to move away from the nucleoside analogs that, until then, had dominated antiviral chemotherapy and toward viral protease inhibitors. This approach promised to disrupt a key step in the viral life cycle—the cleavage of the viral polypeptide chain to produce multiple functional proteins—with potentially less toxicity than conferred by the nucleoside antivirals that inhibit viral genomic replication. Since the introduction of the first HIV protease inhibitors in the early 1990s, proteases have been well-accepted targets for antivirals, and HIV protease inhibitors continue to be part of anti-AIDS cocktails to this day.

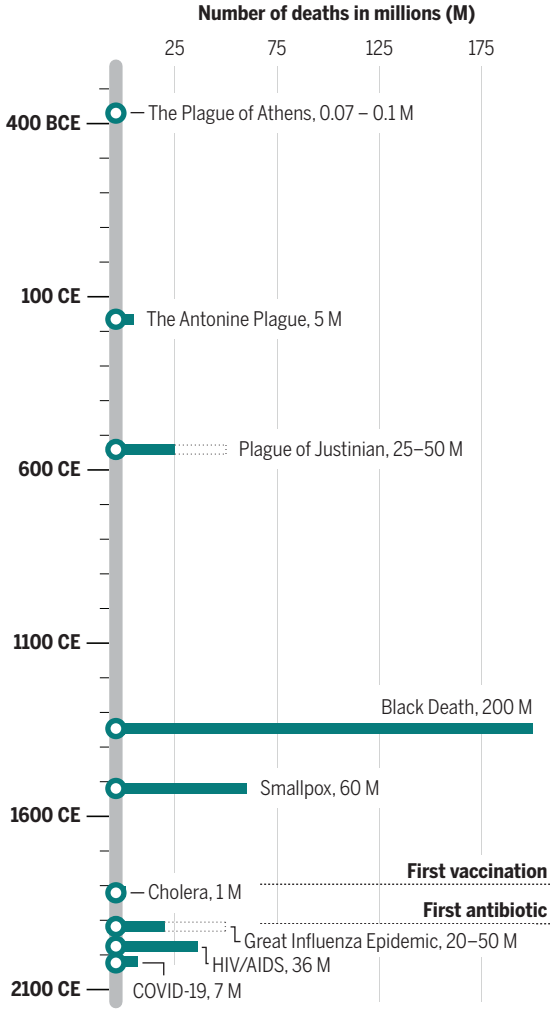
The success of the HIV protease inhibitors inspired the targeting of proteases in hepatitis C virus (HCV), the dengue protease, and the SARS-CoV-2 Mpro and papain-like protease (Plpro). Each has its own story; for example, the intensive combined efforts of academic and industrial scientists to identify a treatment for AIDS targeted HIV protease (PR) by using its three-dimensional structure to computationally screen chemical compounds to rapidly identify an inhibitor (*7*). This is one of the earliest examples of this technique that is now widespread in drug discovery. Notably, and as with Paxlovid, the initial protease inhibitors for HIV PR capitalized on the ~20 years of previous research on renin inhibitors, a human homolog of the viral aspartyl protease. There are now nine US Food and Drug Administration (FDA)–approved drugs against HIV PR. One of them, ritonavir, is used to boost the pharmacokinetic properties of the others, as it does for nirmatrelvir in Paxlovid—another example of borrowing from the past.

The lead therapeutics described by Boby *et al.* may not be ready in time to affect the current pandemic, considering the timelines and challenges of drug approval. Nevertheless, the compounds, and the techniques used to identify them, may well affect human health in the future. The leads themselves represent new departures for Mpro inhibitors, and the methodology used to find them, not to mention the organization of the project, will inspire others. More broadly, the COVID-19 pandemic led to the investigation of several other SARS-CoV-2 enzymes that, like HIV PR in the mid-1980s, have few precedents as antiviral targets but are known to drug discovery more broadly. These enzymes, including those that combat the cellular innate immune system, fool the cell into treating viral RNA as human RNA, and unwind viral RNA to support replication, are crucial for the life cycle of SARS-CoV-2 and are likely good targets for many other viruses as well. Against each of these classes of enzymes, new inhibitors have been discovered that show promise as drug leads (*8*–*11*).

The new classes of antiviral drug leads, of which those reported by Boby *et al.* are among the most advanced, are being developed in both pharmaceutical companies and in universities, including in nine antiviral centers supported by the US government. These efforts seek to restock the antiviral armamentarium for the next pandemic. Such waves of infectious disease have swept through the human population not only in the past century (i.e., influenza, polio, AIDS, and COVID-19) but regularly throughout recorded history. From the plague of Athens (430 BCE), to that of the Antoinines (160 to 189 CE), to that of Justinian (beginning in 541 CE), to the Black Death (14th through 16th centuries), to the smallpox and cholera epidemics of the 17th through the 19th centuries, viral and bacterial plagues have devastated human populations (see the figure). There is every reason to expect these plague cycles to continue (*12*). Until the 1930s, there were no effective drugs to treat microbial pandemics, and, even in recent years, previously discovered drugs were initially relied on to target new threats. The drugs in which investments are made today will pay dividends in the inevitable pandemics of the future. ■

Pandemics from 500 BCE to the present

Pandemics have affected human lives throughout history. Up to 100,000 lives were claimed by the Plague of Athens (*13*) followed by 5 million deaths in the Antonine Plague and ~20 to 50 million deaths in the Plague of Justinian (*14*). The Black Death (*14*) and smallpox (*15*) caused considerable death tolls, and the cholera crisis took 1 million lives (*14*). The advent of the smallpox vaccine in 1796 and the first antibiotic, Salvarsan, in 1910 mark the transition into the modern era of therapeutics against infectious diseases. Vaccines and antibiotics supported the resistance to the Great Influenza Epidemic (*14*), the HIV/AIDS pandemic (*14*), and most recently, COVID-19 (*16*).



REFERENCES AND NOTES

1. D. Lowe, "Covid-19 Small Molecule Therapies Reviewed," *In The Pipeline* (AAAS, 2020); www.science.org/content/blog-post/covid-19-small-molecule-therapies-reviewed.
2. A. Edwards, *J. Chem. Inf. Model.* **60**, 5727 (2020).
3. T. A. Tummino *et al.*, *Science* **373**, 541 (2021).
4. M. L. Boby *et al.*, *Science* **382**, eabo7201 (2023).
5. Y. Duan *et al.*, *Nature* **622**, 376 (2023).
6. A. Douangamath *et al.*, *Nat. Commun.* **11**, 5047 (2020).
7. R. L. DesJarlais *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6644 (1990).
8. T. Otava *et al.*, *ACS Infect. Dis.* **7**, 2214 (2021).
9. C.-H. Zhang *et al.*, *ACS Med. Chem. Lett.* **12**, 1325 (2021).
10. A. Luttens *et al.*, *J. Am. Chem. Soc.* **144**, 2905 (2022).
11. S. Gahbauer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2212931120 (2023).
12. C. Kenny, *The Plague Cycle: The Unending War Between Humanity and Infectious Disease* (Scribner, 2021).
13. R. J. Littman, *Mt. Sinai J. Med.* **76**, 456 (2009).
14. N. Srivastava *et al.*, in *Coronavirus Disease 2019 (COVID-19): Epidemiology, Pathogenesis, Diagnosis, and Therapeutics*, S. K. Saxena, Ed. (Springer, 2020), pp. 9–21.
15. A. Chaikhouni, *Heart Views* **21**, 235 (2020).
16. P. Jha *et al.*, *Lancet* **399**, 1937 (2022).

ACKNOWLEDGMENTS

The authors acknowledge A. L. Lourenço for his artwork. B.K.S. and C.S.C. are partly supported by National Institutes of Health grant U19AI171110.

10.1126/science.adk5868

RETROSPECTIVE

Ian Wilmut (1944–2023)

Embryologist who cloned Dolly the sheep

By **Alan Colman**

Sir Ian Wilmut, the pioneering embryologist who first cloned a mammal from an adult cell, died on 10 September. He was 79. Ian successfully cloned Dolly the sheep using somatic (i.e., postembryonic) cell nuclear transfer (SCNT) with a cell obtained from an adult mammary gland. His achievement, heralded as a scientific wonder, sparked intense ethical debate.

Born on 7 July 1944 in Warwickshire, England, Ian received a BS in agriculture from the University of Nottingham in 1967 and a PhD, for which he researched the preservation of boar semen under the supervision of Chris Polge, from the University of Cambridge in 1971. In 1973, he joined the Animal Breeding Research Organisation in Roslin, Scotland (now named the Roslin Institute). He became a professor of reproductive science and director of the new Centre for Regenerative Medicine (in 2005 and 2006, respectively) at the University of Edinburgh. He retired in 2012 but remained active in science until health issues forced him to step back in 2018.

Ian's early work at the University of Cambridge focused on freezing, thawing, and implanting cow embryos and led to the birth of Frostie, the first calf born from the implantation of a frozen embryo. In 1982, he joined the team that—in 1985—produced the first transgenic sheep that secreted human therapeutic proteins into its milk. The process of making a transgenic sheep at that time was inefficient; only 1 to 3% of live births contained the transgene.

Ian wanted to improve transgenesis efficiency and also to make precise changes to the genomes of large animals, an objective already achieved in mice using embryonic stem (ES) cell technology. No one, Ian included, had succeeded in making domestic animal ES cells, so Ian turned his attention to improving nuclear transfer in sheep. Mammalian nuclear transfer (cloning) had all but stalled, but in 1986, Steen Willadsen cloned sheep from 4-day-old embryo cell nuclei. In 1989, Ian and his PhD student Lawrence Smith reproduced Willadsen's experiments and reported the birth of four live sheep.

Ian appreciated that to use cloning for genetic manipulation, he would need large, cloning-competent cell populations. He thought that, although later-stage embryo-derived cells might suffice initially, he would eventually need to use somatic cells. With characteristic insight, he reasoned that synchronizing the cell cycle status of the recipient egg cytoplasm and the incoming nucleus would be essential. In 1991, he recruited cell biologist Keith Campbell to solve this challenge. Keith worked out the ideal (quiescent) cell donor and recipient and deliberately triggered the development of the reconstructed embryo by electrical activation at the same time as the nuclear transfer. In 1995, Ian and Keith used this novel protocol with cells cultured from a 9-day-old sheep embryo. Two cloned lambs, Megan and Morag, were born in 1996. The practical implications of this tour de force prompted Pharmaceutical Proteins Ltd. (PPL)—an independent company established in 1987 to exploit Roslin's transgenic technology—to propose a collaboration with the institute to validate two PPL embryonic cell lines and a somatic cell line (fetal fibroblast) suggested by Ian. When the second PPL line became unusable, PPL's Angelika Schnieke proposed an adult (somatic) cell line cultivated from the frozen mammary tissue of a deceased 6-year-old Finn Dorset ewe. The prospect of success seemed very poor to everyone, including Ian, who was well aware of the failure of SCNT pioneer Sir John Gurdon to produce adult frogs using adult cell donors.

Yet Dolly was born on 5 July 1996, the only live birth from 277 implanted embryos. The embryonic line produced four lambs, the fetal line produced two. Dolly was perfectly formed (although the cover of *Nature* erroneously indicated otherwise, to Ian's amusement). Ian quickly countered criticism alleging that Dolly could not have been cloned from an adult cell nucleus and, with other collaborators, provided additional scientific corroboration. Meanwhile, he found it hilarious that a UK patent examiner needed no such convincing before rejecting a patent for Dolly on the grounds that her copy had existed in the wild.

Within weeks of the Dolly publication, Ian and the team were summoned to the UK Parliament to explain the work's implications for humans. Ian spoke optimistically but warned the panel that humans could potentially be cloned within a year, a development he feared.

Over the next 3 years, others cloned mice and cows from adult cells, while Ian's team and PPL vindicated his vision for using SCNT to improve the genetic manipulation of livestock. In 1997, Polly and Molly, transgenic sheep expressing human coagulation factor IX, were born, and in 2000, Cupid and Diana resulted from the successful targeting of the sheep collagen locus. These milestones used the type of cell that Ian favored: genetically manipulated fetal fibroblasts.

Ian's many accolades included his election to the Royal Society in 2002 and the National Academy of Sciences in 2004, the Paul Ehrlich and Ludwig Darmstaedter Prize in 2005, and the Shaw Prize in Life Science and Medicine and a knighthood in 2008. He used his celebrity status to highlight the dangers of human cloning. He also pushed for the development of motor neuron disease models using patient-specific pluripotent stem cells, which were made with SCNT until the development of human induced pluripotent stem cells (iPSCs) in 2007.

I met Ian in 1987, in my capacity as research director of PPL. Affable and soft-spoken, he had an avuncular manner and dry sense of humor that served him and the scientific community well. His friends and neighbors knew him for his passion for hill walking, curling, and the occasional tot of whiskey.

Ian pioneered the use of SCNT in mammals, and Dolly's birth was an unexpected bonus that dispelled the dogma that an isolated adult vertebrate nucleus could not be reprogrammed by SCNT back to a pristine embryonic state. Shinya Yamanaka, cowinner (with John Gurdon) of the 2012 Nobel Prize in Physiology or Medicine, acknowledged that Dolly inspired his successful derivation of iPSCs in 2006. Ian was rightfully proud of his contribution. ■

10.1126/science.adl1593

Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA.
Email: colman@fas.harvard.edu

PHOTO: WATTIE CHEUNG/CAMERA PRESS/REDUX

INSIGHTS

A house burns in Malibu, California, on 9 November 2018 during the Woolsey Fire.

POLICY FORUM

ENVIRONMENTAL ECONOMICS

Adapting to growing wildfire property risk

Cost-effective and equitable physical and financial protections are needed

By **Judson Boomhower**[1,2]

Wildfire-threatened communities are on the front lines of climate change. From 2013 to 2022, the share of global disaster losses caused by wildfires more than doubled compared with losses in previous decades (*1*). On page 702 of this issue, Radeloff *et al.* (*2*) draw on a 30-year time series of housing counts and vegetation to show how housing expansion, area burned, and vegetative fuels contribute to wildfire losses and the increasing number of homes at risk in the United States. With tens of millions of US homes now confronting a growing risk of destruction by wildfires, adaptation is an urgent policy and research challenge. Success will require scaling up cost-effective investments in physical protection to reduce wildfire losses, ensuring well-functioning insurance markets to absorb risk that cannot be cost-effectively mitigated away, and addressing disparities in protection and postfire recovery for socially vulnerable populations.

INSIGHTS FROM THE NEW WORK

Radeloff *et al.* offer several scientific insights. One contribution is decomposing the growth in homes that are actually experiencing wildfires. Have people moved to the risk, or is the risk moving to people? The answer, they found, is both. Between the 1990s and 2010s, 47% of the increase in homes that experience wildfires was due to housing expansion, whereas 53% was due to increased wildfire occurrence.

A second set of results focuses on the number of homes at risk of wildfire. Following the forestry literature and their own pioneering work, Radeloff *et al.* focus on a discrete high-hazard zone based on federal land-management criteria (*3*, *4*). Census blocks are classified as part of this wildland-urban interface (WUI) if they exceed certain population and vegetation thresholds. The number of homes in the

WUI increased by a staggering 46% between 1990 and 2020, reaching 44 million. However, the most recent data compiled by Radeloff *et al.* show a slowing in this growth rate. About half as many new WUI homes were built during 2010 to 2020 compared with 1990 to 2000. Moreover, many of these recent homes were built in already-developed WUI areas, so that the total areal footprint of the WUI expanded very little. In addition, the composition of new construction has shifted: During 2010 to 2020, 30% of all new US homes were built in the WUI, down from 39% in previous decades. There is important regional variation in these trends; for example, WUI growth in Texas accelerated according to the most recent data.

The third insight from Radeloff *et al.* is about grassland and shrubland vegetation. The large extent of these fuel types in the western United States means that grassland fires destroy more total homes than forest fires, even though forest fuel types are associated with higher risk of property loss in a given fire.

This 30-year time series of wildfire-prone development based on a consistent methodology is valuable for research and policy design. Understanding how and where the WUI is growing or shrinking is crucial for designing and evaluating policies. It also allows researchers to unpack topics such as the effects of a broader housing market slowdown on WUI growth, as Radeloff *et al.* do. Finally, the long data series enriches understanding of the correlates of risk and loss, such as the results on vegetation types that the authors present.

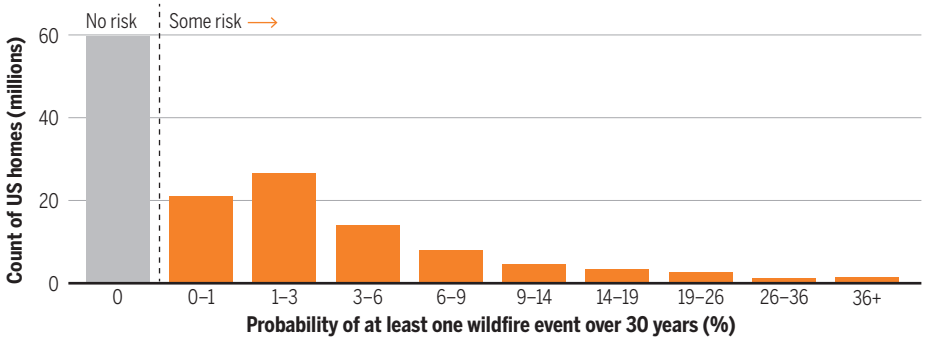
Conversely, applying a consistent methodology over many years restricts the ability to incorporate new advances. The WUI classification is valuable as a national- or regional-scale measure of long-term trends, but it has limitations as a proxy for wildfire hazard to a given community or property. The binary "interface" and "intermix" labels do not capture important continuous variation in wildfire hazard. Census blocks, the basic unit of WUI geography, can be large, which further obscures local variation in risk. And the classification thresholds—for example, at least one home per 40 acres, wildland vegetation covering 50% of a census block—are somewhat arbitrary.

Fortunately, there has been tremendous innovation in spatially granular models of wildfire hazard and risk, including models from federal agencies (*5*), nonprofits, and universities (*6*), and proprietary models developed by firms like AIR, CoreLogic, and Risk Management Solutions. Radeloff *et al.* also embrace this trend toward granularity when considering the association between vegetation and structure-level damage outcomes.

The rest of this Policy Forum discusses the implications of Radeloff *et al.* and recent related work for wildfire policy. Radeloff *et al.* enumerate a variety of possible policy responses to wildfire risk, from building standards to changes in the electric grid to Smokey the Bear. Identifying specific policies that can most efficiently promote adaptation is a crucial area for research. Because the authors focus on threats to built structures, the discussion below is similarly bounded. A complete wildfire policy agenda will also need to address air pollution, ecosystem and watershed impacts, health risks for first responders, and other issues (*7*).

Distribution of wildfire probabilities for US homes

The chart is based on data from the First Street Foundation's Fire Factor model. Bins include lower bounds and exclude upper bounds, except for 0 to 1%, which excludes both the lower and upper bounds. See supplementary materials for data-source information.



PUTTING NUMBERS TO RISKS

For what follows, it will help to enrich the binary risk categorization in Radeloff *et al.* with something more granular. The Fire Factor model produced by the nonprofit First Street Foundation predicts wildfire hazard for all US homes (*6*). This model, one of several frontier models, has the advantage of being a relatively accessible open-science model.

The First Street Fire Factor model estimates the distribution of wildfire likelihood across 143 million US homes (see the figure). Of those homes, 83 million face nonzero wildfire risk. For most of these threatened homes, wildfire exposure is a rare event. The probability of experiencing one or more wildfires over 30 years is 3% or less for 48 million of the 83 million threatened houses. That corresponds to an annual wildfire probability of 0.1% or less. At the high end of the risk spectrum, 5 million homes face a 30-year event probability greater than 19%.

In comparison to flooding, another climate-related risk, wildfires threaten a larger number of homes, but the hazard faced by threatened homes is smaller on average. In a 2020 study, the First Street Foundation found that 22 million US homes faced risk from flooding, about one-fourth the number threatened by wildfires. However, 67% of flood-threatened homes have 30-year flood probabilities greater than 26%, and 17% are virtually guaranteed to flood (30-year event probability greater than 99%) (*8*).

The sheer number of wildfire-threatened homes means that many households face a nontrivial chance of a terrible loss. The fact that most of these threatened homes have relatively low event probabilities has two implications for adaptation. First, it affects the return on adaptation investments. Second, the fact that many people face a small probability of a big loss makes insurance an especially important part of successful adaptation. The following two sections discuss these implications in turn.

DRIVING UPTAKE OF HIGH-RETURN ADAPTATION INVESTMENTS

Many of the policies mentioned by Radeloff *et al.* and others aim to increase protective investments like "structure hardening" (using fire-resistant building materials), maintaining defensible space around buildings, and implementing community-scale vegetation management. The wildfire hazard distribution underscores how important it will be to prioritize cost-effectiveness as these efforts are scaled up (see the figure).

A simple cost-benefit comparison helps to elucidate the importance of focusing on high-return investments. Consider this crude proxy for the net benefits of

[1]Department of Economics, University of California, San Diego, La Jolla, CA, USA. [2]National Bureau of Economic Research, Cambridge, MA, USA. Email: jboomhower@ucsd.edu

GRAPHIC: D. AN-PHAM/*SCIENCE*

hardening a single home. The reduction in expected financial losses is the product of the annual wildfire probability p, the value of the potential loss V, and the change in structure-survival probability that is conditional on a wildfire occurring Δq. A simplified investment rule compares the upfront cost c to the flow of avoided losses over the life of the home, using stylized discount factor δ. The resulting cost-benefit test is $\delta pV\Delta q \geq c$.

The low values of p for many wildfire-threatened homes mean that investments must have low costs and/or large effects on survivability to pass this cost-benefit test. The good news is that while a new house is being built, there are a variety of home hardening investments with modest costs and large survivability benefits that are available for use, such as fire-resistant roofing materials, siding, eaves, gutters, and vents. These investments can effectively and cost-effectively reduce risk for new buildings (*9*, *10*).

The not-so-good news is that there are fewer promising hardening options for the millions of already-built wildfire-threatened homes. The cost of fully retrofitting an existing home to modern standards is tens of thousands of dollars (*10*). Even with generous values for V and Δq of $1.5 million and 25 percentage points and an annual discount rate of 5%, an investment with a $65,000 upfront cost and a 30-year lifetime would be cost-effective for homes where p exceeds 1.1%. A 1.1% annual probability implies a 30-year event probability of 28%. Only about 3% of wildfire-threatened homes face such an extreme exposure probability (see the figure).

Of course, this simple "expected breakeven" criterion ignores important details. More-sophisticated calculations can consider risk aversion and ancillary benefits of protective investments such as decreased public expenditures on emergency response (*9*). Even after accounting for these additional benefits, the economics of hardening already-built homes remain, for the most part, unappealing. Given these facts, policymakers in California and elsewhere have emphasized low-cost investments for existing homes such as managing vegetation and other fuels around the home to create defensible space and placement of fire-resistant screens on vents to prevent embers from entering the home (*11*).

In addition to these parcel-level investments, a range of community-level investments could reduce wildfire risk to homes. These include vegetation management for shared spaces and public lands, additions to firefighting capacity, and education campaigns to reduce accidental ignitions. Measuring the effectiveness and cost-effectiveness of these community-scale adaptations is an important research problem.

Identifying cost-effective physical adaptations is one thing. What role should policy-makers have in implementing these investments? Adaptation efforts confront a plethora of potential market frictions, including misperception of risk by property owners, spatial externalities (one person's decisions affect the risk faced by their neighbors), insurance prices that may not fully reward protective investment, and incentive effects created by disaster-relief policies (*12*). Researchers must identify which of these market barriers most pressingly constrain adaptation. Policy-makers can target these barriers with appropriate remedies such as building codes or investment subsidies. Rigorous evaluation of these interventions using randomized control trials or quasi-experimental methods will be critical to identify successful policies.

As Radeloff *et al.* observe, it is important to consider not only how homes are built but also where. Incentive conflicts in the permitting of new construction may lead to excess growth in areas of high wildfire risk, with public expenditures on wildfire protection subsidizing development in harm's way (*13*). At the same time, long return intervals (see the figure) and the large sunk investments in already-built homes mean that managed retreat—the strategic abandonment of assets at risk of disaster—has less of a role to play in areas with wildfire risk than in areas facing permanent inundation by rising seas or frequent repetitive flooding.

THE IMPORTANCE OF INSURANCE

Economically feasible investments in property protection cannot fully counteract the growing threat from wildfires, especially for the large stock of older homes. The second critical adaptation lever is to improve the ability of insurance markets to provide financial protection when physical defenses fail. Given this, it is concerning to see wildfire risk roiling property insurance markets. Even as prices have increased, the availability of coverage has contracted, and participation in state-run residual plans, such as the California FAIR plan, has grown. There is an urgent need for economic research to understand these market challenges.

Wildfires and other natural catastrophes challenge insurance markets in several ways. First, the rare nature of these events means that there is too little historical data for insurers to confidently price the risk using traditional actuarial methods. Stochastic catastrophe models of wildfire risk are starting to fill this gap, as they have for hurricane and earthquake risk. In addition to helping improve the models, there may be a role for policy-makers to

A house near Cle Elum, Washington is surrounded by burnt terrain after surviving a wildfire on 14 August 2012.

lessen barriers to adoption. Research that is in progress documents surprising disparities in the sophistication of wildfire-risk ratings across major property insurers in California.

In addition to better models of wildfire ignition and spread, there is an opportunity to improve data on property characteristics and investments in risk reduction. For insurance contracts to properly reward protective investments, insurers must be able to observe these investments at low cost. Centralized databases of property characteristics and/or trusted certifications from groups such as the Insurance Institute for Business and Home Safety will reduce the need for redundant monitoring investments by each insurer.

The spatially correlated nature of disaster risk exposes insurers to potentially huge payouts, which requires them to hold costly capital surplus or purchase reinsurance to protect their solvency. This issue of "risk load" has long been central in markets for hurricane risk. It will also play a key role in the "new normal" of insuring wildfire risk. Responding to this challenge will require better modeling of tail events combined with financial innovations to transfer tail risk through reinsurance, insurance-linked securities, carefully designed public-private partnerships, and other tools (*12*).

Finally, property insurance markets in the United States feature considerable regulation. In an environment of increasing risk, regulations meant to protect the affordability of insurance—including limits on the pace of price increases and restrictions on what information may be used to justify prices—may inadvertently affect firms' incentives to offer coverage. As wildfire risk grows, state regulators must ensure that well-intended regulations do not prevent necessary upward adjustments in insurance prices.

EQUITABLE WILDFIRE ADAPTATION

Vulnerability to wildfires depends not only on physical fire risk but also on demographic and economic factors. Existing research finds wide heterogeneity in indicators of social vulnerability within wildfire-hazard areas. On average, homeowners in high-hazard zones appear to be slightly wealthier than those in surrounding areas. At the same time, certain socioeconomic groups that may be more vulnerable—for example, the very poor, the elderly, and Native Americans—are overrepresented in wildfire areas (*14*). There also appear to be disparities in access to protective investments. This is true not only for costly private investments but also for public programs: Wealthier, whiter communities have been more successful in lobbying for federal fuel-mitigation projects (*15*). An important challenge for adaptation policy will be to find a way to target financial support for protective investment where it is most needed.

Households also differ in their ability to recover from a wildfire loss. Most homeowners have homeowners insurance, but it is common for the value of losses to exceed policy limits. Renters may face a complete loss of their belongings. Factors such as education and income may also influence an insured individual's ability to enforce the terms of the insurance contract. Regulators can require insurers to report claim denial rates, processing times, and related statistics at a sufficiently granular scale to monitor such potential disparities. Finally, some types of losses are difficult or impossible to insure. Economically vulnerable families that experience a wildfire disaster may face long-lasting disruptions to employment, school attendance, and social connections. Research is needed to better understand these long-term effects and strategies to reduce them.

IDENTIFY, EVALUATE, AND SCALE

The new data reported by Radeloff *et al.* underscore the need for well-designed policies to guide adaptation. One priority will be identifying and scaling up cost-effective physical investments that can reduce wildfire losses. As new WUI construction appears to slow and technology for hardening homes at the time of construction continues to improve, the toughest challenge will be to efficiently protect the large stock of older homes in high-hazard areas that were built with little thought to wildfire. A second priority is strengthening insurance markets to transfer remaining risk. Finally, policymakers will need to ensure that policies for ex ante protection and ex post recovery address the needs of low-income, minority, and other socially vulnerable communities. All of these goals will benefit from new and emerging spatially granular models of wildfire risk.

Decades of careful research like that presented by Radeloff *et al.* have built a strong scientific foundation to inform adaptation. Social scientists and policy-makers must build on this foundation to identify promising market and policy instruments, rigorously evaluate these prototypes, and scale the most effective solutions. ■

REFERENCES AND NOTES

1. C. Banerjee, L. Bevere, T. Corti, J. Finucane, R. Lechner, "A perfect storm: Natural catastrophes and inflation in 2022" (Swiss Re Institute, 2023).
2. V. C. Radeloff *et al.*, *Science* **382**, 702 (2023).
3. V. C. Radeloff *et al.*, *Ecol. Appl.* **15**, 799 (2005).
4. US Department of Agriculture, US Department of the Interior, *Fed. Regist.* **66**, 751 (2001).
5. J. H. Scott *et al.*, Wildfire risk to communities: Spatial datasets of landscape-wide wildfire risk components for the United States. Forest Service Research Data Archive (2020).
6. First Street Foundation, Risk Factor (2023); https://firststreet.org/risk-factor/.
7. T. J. Feo, S. Evans, A. J. Mace, S. E. Brady, B. Lindsey, "The costs of wildfire in California: An independent review of scientific and technical information" (California Council on Science and Technology, 2020).
8. First Street Foundation, "The first national flood risk assessment: Defining America's growing risk" (First Street Foundation, 2020).
9. P. W. Baylis, J. Boomhower, "Mandated vs. voluntary adaptation to natural disasters: The case of U.S. wildfires," Working paper 29621, National Bureau of Economic Research, Cambridge, MA, December 2021.
10. Headwaters Economics, "Building a wildfire-resistant home: Codes and costs" (2018).
11. CAL FIRE, Low-cost retrofit list (2022); https://osfm.fire.ca.gov/media/p0elt0sp/low-cost-retrofit-list-update-2_17_22.pdf.
12. H. C. Kunreuther, E. O. Michel-Kerjan, *At War with the Weather: Managing Large-Scale Risks in a New Era of Catastrophes* (MIT Press, 2011).
13. P. Baylis, J. Boomhower, *Am. Econ. J. Appl. Econ.* **15**, 442 (2023).
14. M. Wibbenmeyer, M. Robertson, *Environ. Res. Lett.* **17**, 064031 (2022).
15. S. Anderson, A. J. Plantinga, M. Wibbenmeyer, *J. Polit.* **85**, 625 (2022).

ACKNOWLEDGMENTS

J.B. acknowledges generous funding from the Carnegie Corporation of New York through an Andrew Carnegie Fellowship. P. Baylis, M. Fowlie, J. Gellman, and A. Plantinga provided helpful discussions.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adk7118

10.1126/science.adk7118

A pair of great crested grebes (*Podiceps cristatus*) perform a courtship ritual.

BOOKS *et al.*

BIOLOGY

Birds of a feather

An ornithologist mobilizes feminist scholarship to advance a holistic view of gender

By **Erika Lorraine Milam**

Richard Prum is an evolutionary biologist by training and an ornithologist by calling. To meet him in the field or in the library is to be caught up in the passion with which he engages the world, whether he is relating his latest bird adventure or discussing the most recent paper he has read. As is evident from his most recent book—*Performance All the Way Down*—he has been doing a great deal of reading indeed.

At stake for Prum is a particular understanding of what it means to be human. "We are *not* products of lawlike, genetic determination mechanisms," he avers, "but iterative, individualized bodily self-enactments situated in specific biological and social environments." Readers of his earlier work, including his 2017 book, *The Evolution of Beauty*, will find themselves intrigued by his continued engagement with feminist science studies—and he has done his homework.

Performance All the Way Down: Genes, Development, and Sexual Difference
Richard O. Prum
University of Chicago Press, 2023. 368 pp.

In *Performance All the Way Down*, Prum extends the notion of biological agency to the molecular processes and developmental pathways that create sexual variation in the natural world. The result is a nonbinary understanding of sexual expression, as characteristic of animals and plants as it is of humans. Sex, Prum posits, is not a static state of being or a neutral epistemological category. Instead, sex arises from the enactment of the self at every level of biological organization, from molecules to organisms.

Throughout the book, Prum emphasizes the interrelationality of developmental processes and performative pathways, the molecular discourse that underpins them, and the gender "citationality"—the tendency of individual gender expression to reference current and previous enactments of gender—that emerges. Prum's vision of agency in the living world extends his analysis in *The Evolution of Beauty* to the interactions required of molecules in developmental pathways that respond to the environmental and social conditions in which an individual is situated.

Performance All the Way Down is replete with citations to the rich research on this topic put forth by feminist scholars such as Karen Barad, Donna Haraway, Judith Butler, Helen Longino, Sarah Richardson, Anne Fausto-Sterling, Elizabeth Grosz, Patricia Gowaty, Malin Ah-King, Eve Sedgwick, Lily Kay, Rebecca Jordan-Young, Katrina Karkazis, and Banu Subramaniam. Prum mobilizes this work to conceptualize biological processes as inherently agential, interactive, and (his ultimate quarry) performative.

Citations are, of course, both acts of acknowledgment and political processes, which is why humanists often emphasize citational practices. Framing the act of citation as an interrelational practice allows these scholars to highlight the social and collaborative processes that constitute the production of scholarship. Read this way, Prum's own citational practices echo his explanation of how sex is produced in organic bodies—both are relational, and both reflect the diversity of their origins. His book exemplifies these self-reflective practices, from introduction to index.

Prum is not the first biologist to have reached for the tools of the humanities to better communicate and convince readers of their scientific arguments. An obvious comparator is Joan Roughgarden, who drew more lightly from humanistic perspectives in *Evolution's Rainbow: Diversity, Gender, and Sexuality in Nature and People* (2004) to provide her own biological explanation of the origins of sexual diversity in nature. A slightly more distant but equally compelling comparator is the venerable Ernst Mayr, who back in the 1980s mobilized the history of science to shore up his vision of biology as divided into two almost incommensurable approaches to understanding the transformations inherent to life on this planet: a reductionist, genetic perspective and an integrative, evolutionary one. (Mayr, it need not be said, was a fan of the latter.) That natural scientists continue to recognize historians, sociologists, and anthropologists as intellectual allies means we have come a long way since the contentious debates about scientific authority and fallibility that defined the science wars of the 1990s.

Performance All the Way Down contains a lot of big ideas, both because of the biological content Prum strives to convey to his readers and because of the sophisticated nature of the feminist theory he mobilizes. This review is a necessarily short gloss on this thoughtful and ambitious book. To understand the full scope of Prum's vision, I encourage you to read it in full. ■

The reviewer is at the Department of History, Princeton University, Princeton, NJ 08544, USA.
Email: emilam@princeton.edu

10.1126/science.adj1203

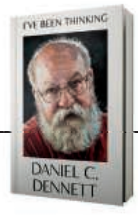
PHILOSOPHY

Thinking about science

A philosopher reflects on his influential interrogations of free will, consciousness, and artificial intelligence

By **Itai Yanai**[1] *and* **Martin J. Lercher**[2]

I've Been Thinking
Daniel C. Dennett
Norton, 2023. 464 pp.



One could call Daniel Dennett the "scientists' philosopher," as for more than six decades his work has involved close collaborations with scientists. A common theme is woven through his impressive body of work: He begins by setting his sights on a seemingly magical or mysterious concept—consciousness, for example, or free will—and then proceeds to systematically dissect it using thought experiments and empirical evidence. In his new memoir, *I've Been Thinking*, Dennett not only tells the story of his extraordinarily broad academic and intellectual life but also explains the tools he uses to advance his thinking. These include a tactic Dennett has referred to in past works as an "intuition pump," in which an audience is led to a specific conclusion by means of intuitive logic (*1*).

Dennett's intellectual process thrives on a balance between discussions with fellow philosophers and scientists and a solitary process his friend Douglas Hofstadter once called "tillosophy," in which Dennett begins by confronting a hard problem and persists until he realizes that he is stuck, at which point he deliberately engages in activities that are less intellectually demanding. These latter activities often include tilling one of the fields on his Maine hobby farm. This practice frees his brain to mull over the philosophical problem with which he is grappling, helping his thinking to move forward, while at the same time ensuring progress is made on the farm's many chores.

Dennett's favorite activity—centered at Tufts University for most of the past 52 years—seems to be engaging in conversation with scientists. He revels in discussions with his ever-evolving circle of scientific colleagues, and much of the book is devoted to recounting how such interactions have shaped his thinking. The crucial role played by his collaborations with scientists is most evident in the development of the ideas Dennett explored in his influential 1991 book, *Consciousness Explained* (*2*), which prompted many scientists to embark on the testing of its ideas and also to perform their own philosophizing on the topic of consciousness.

In addition to consciousness, Dennett has worked on two other fundamental phenomena: free will and the appearance of design in evolution by natural selection. Unfortunately, *I've Been Thinking* gives much less detail about his important contributions to these fields, with the exception of recounting his most memorable feuds with the eminent evolutionary biologist Stephen Jay Gould.

Dennett mans the helm of his sailboat, the *Xanthippe*.

PHOTO: DAN DENNETT FAMILY

Beyond biology and cognitive science, Dennett has repeatedly turned to computer science for inspiration on how the mind works. His engagement with this field began with a visiting professorship at Harvard in 1974, when the topic of interest was what is now referred to as "Good-Old-Fashioned AI" (or GOFAI)—artificial intelligence (AI) based on high-level, human-readable representations of knowledge. Thinking about how expert systems can be built from simple logical statements helped Dennett explain how mental processes of astonishing complexity can arise from much simpler building blocks.

More recently, the advent of neural networks has led to powerful artificial systems that more closely mimic the information processing performed by our brains. Dennett is both in awe of and deeply concerned by recent developments in this field. He describes how an AI version of himself—"DigiDan"—was created by training OpenAI's GPT-3 with his published work, and how he was amazed to find that even his closest colleagues could not consistently distinguish between statements written by him and those produced by the synthetic philosopher. In a postscript, he describes his concern about the imminent threat of "counterfeit people" generated by large language models such as ChatGPT, which exploit our natural instinct to treat conversational partners as rational agents.

"At its best, philosophy is intellectual reverse engineering, methodically dismantling bad habits of thought," writes Dennett. His philosophy is admirably inclusive, appealing to nonexpert audiences with accessible language. He is also acutely aware that this approach has a second benefit: By ostensibly writing for the public, he is also offering his fellow philosophers and scientists an easy-to-follow version of his arguments without the risk of offending them by overexplaining.

One shortcoming of *I've Been Thinking* is its disproportionate focus on Dennett's earlier work. Readers learn very little about the conception of his bestselling book *Darwin's Dangerous Idea* or his more recent book, *From Bacteria to Bach and Back*, for example. But perhaps it is a good sign that the reader is left wanting more of Dennett's stories. It leaves open the possibility of a sequel. ■

The reviewers are at the [1]Institute for Systems Genetics, NYU Langone Health, New York, NY 10016, USA, and [2]Department of Biology and Institute for Computer Science, Heinrich Heine University, 40225 Düsseldorf, Germany.
Email: itai.yanai@nyulangone.org

REFERENCES AND NOTES

1. D. C. Dennett, *Intuition Pumps and Other Tools for Thinking* (Norton, 2013).
2. D. C. Dennett, *Consciousness Explained* (Back Bay Books, 1991).

10.1126/science.adk5070

Panama's mangrove forests can help the country combat the effects of climate change.

Edited by **Jennifer Sills**

Panama must protect mangroves and peatlands

Mangroves and peatlands in Panama play a vital role in preserving biodiversity and provide critical ecosystem services in Central America. However, despite their importance, Panama has yet to fully enshrine mangroves or peatlands in its climate strategy. Given their far-reaching benefits, Panama should improve protection for mangroves and peatlands.

Located within two biodiversity hotspots (*1*), Panamanian mangroves and peatlands are part of the Mesoamerican Biological Corridor and are integral to the main transcontinental American flyways, securing the annual migration of millions of birds (*2*, *3*). Nature-based solutions to climate change, these ecosystems store a substantial portion of Central America's irrecoverable carbon, which is both vulnerable to release and challenging to restore (*4*). Mangroves and peatlands have also been shown to mitigate greenhouse gas emissions (*5*, *6*), increase ecosystem resilience, and conserve biodiversity (*7*, *8*).

In October, Panama hosted the 2023 Latin America and the Caribbean Climate Week (*9*), strengthening the country's role as a leading regional climate change advocate. Panama is also engaged in a nationwide public consultation process (*10*) to shape its plan to combat climate change and meet greenhouse gas emissions reduction targets. This process presents a valuable opportunity to enhance Panama's climate strategy by integrating more ambitious pledges for mangroves protection and restoration, and for the first time including peatland conservation in its goals.

Moving forward, Panama should enhance its protection of carbon-rich ecosystems by improving mangroves and peatlands inventories. Identifying these important ecosystems will provide a reference for conservation efforts. Panama should also develop accurate monitoring systems to track ecosystem change and restoration initiatives, and it should strengthen community-based governance models to guarantee the sustainability of future conservation efforts. By taking these steps, Panama could improve the effectiveness of its climate change strategy and contribute to addressing the global climate and biodiversity crises.

Jorge Hoyos-Santillan[1,2,3,4]*, Alejandro Miranda[4,5], Ligia Castro de Doens[6], Esperanza González-Mahecha[7], Miguel Cifuentes-Jara[8,9], Thomas A. Worthington[10]

[1]Smithsonian Tropical Research Institute, 0843–03092 Panama City, Panama. [2]Audubon Americas, New York, NY 10014, USA. [3]School of Biosciences, University of Nottingham, Sutton Bonington LE12 5RD, UK. [4]Center for Climate and Resilience Research, 8320000 Santiago, Chile. [5]Laboratorio de Ecología del Paisaje y Conservación, Departamento de Ciencias Forestales, Universidad de La Frontera, 4811230 Temuco, Chile. [6]Dirección de Cambio Climático, Ministerio del Medio Ambiente, 07116 Panama City, Panama. [7]Inter-American Development Bank, 269–6900 Panama City, Panama. [8]Conservation International, Arlington, VA 22202, USA. [9]Centro Agronómico Tropical de Investigación y Enseñanza, 30501 Turrialba, Costa Rica. [10]Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge CB2 3QZ, UK.
*Corresponding author. Email: jorge.hoyos@umag.cl

REFERENCES AND NOTES

1. N. Myers, R. A. Mittermeler, C. G. Mittermeler, G. A. B. Da Fonseca, J. Kent, *Nature* **403**, 853 (2000).
2. M. A. Smith *et al*., *Bird Migration Explorer* (National Audubon Society, New York, NY, 2022).
3. B. King, "Panama shatters raptor migration record," *Smithsonian Insider* (2014).
4. M. L. Noon *et al*., *Nat. Sustain.* **5**, 37 (2021).
5. D. C. Donato *et al*., *Nat. Geosci.* **4**, 293 (2011).
6. J. Leifeld, L. Menichetti, *Nat. Commun.* **9**, 1071 (2018).
7. M. D. Spalding, M. Leal, *The State of World's Mangroves 2022* (Global Mangrove Alliance, 2022).
8. Global Peatlands Initiative, "Global Peatlands Assessment—The State of the World's Peatlands: Evidence for action toward the conservation, restoration, and sustainable management of peatlands" (Nairobi, 2022).
9. UN Climate Change, *Latin America and the Caribbean Climate Week 2023* (2023); https://unfccc.int/LACCW2023.
10. Government of Panama, "El país avanza en el proceso de preparación de su Segunda Contribución Determinada Nacionalmente (CDN2)" (2023); https://www.miambiente.gob.pa/el-pais-avanza-en-el-proceso-de-preparacion-de-su-segunda-contribucion-determinada-nacionalmente-cdn2-2 [in Spanish].

COMPETING INTERESTS

E.G.-M. is funded by the UK Blue Carbon Fund.

10.1126/science.adl3048

Mangrove decline puts Pakistan's coasts at risk

Climate change threatens the coastal areas of Pakistan by increasing sea levels (*1*), soil erosion, and pollution (*2*). Intact mangroves play a crucial role in preserving the ecological balance of coastal regions by absorbing

excess nutrients, stabilizing sediments, and sequestering carbon, thereby improving water quality and mitigating pollution (*3*). However, Pakistan's mangroves have experienced substantial degradation over the past 50 years (*4*). Pakistan should increase its coastal protection measures to guard against its vulnerabilities to climate change.

The effects of climate change in Pakistan's coastal areas are costly. The rise in sea levels and increased storm activity result in the accumulation of soil and mud in harbor channels, necessitating substantial financial investments by harbor authorities for regular cleaning and maintenance (*5*). Erosion and runoff from extreme weather events have led to the spread of toxic heavy metals among marine wildlife and ecosystems (*6*), and such pollutants can infiltrate the human food chain, causing disease and acute health problems (*7*). Rising temperatures can reduce photosynthesis and increase salinity levels, leading to a decrease in the availability of food for microorganisms that are vital components of the marine food web (*8*).

Mangroves can help protect against the effects of climate change. Their intricate root systems trap sediment and slow the flow of water, reducing the risk of coastal erosion and flooding during extreme weather events. They also serve as a natural defense against pollution, filtering pollutants and preventing their entry into coastal waters. Mangroves sequester substantial amounts of carbon dioxide (*9*), helping to combat global warming. In addition, mangroves support fisheries, increasing food security and economic stability in coastal regions. The loss of just 1.6 km^2 of mangrove trees in Pakistan results in 30 to 60% fewer total fish produced per year (*10*).

Pakistan's prioritization of agriculture and industry over coastal ecosystems has led to the decline in its mangrove forests. Freshwater resources have been redirected from natural ecosystems and water bodies to agricultural fields. Industrial and urban activities have contaminated the water on which the mangroves depend. The country has also allowed excessive fishing in its coastal waters, which has had a detrimental impact on marine ecosystems and fish populations (*11*).

Pakistan is highly susceptible to the adverse consequences of sea level rise (*12*), but the country has placed insufficient emphasis on coastal area protection and flood control measures. To protect its mangroves, Pakistan should pass legislation that limits logging, land conversion, and pollution. The government should also engage communities through education about the importance of mangroves and provide job training to help community members reduce local resource dependence by pursuing livelihoods other than fishing. Finally, the country should prioritize restoration by planting native mangrove species in degraded areas to enhance resilience against climate change.

Uzair Aslam Bhatti[1,2], Hao Tang[2], Shulei Wu[1]*
[1]School of Information Science and Technology, Hainan Normal University, Haikou 571158, China. [2]School of School of Information and Communication Engineering, Hainan University, Haikou 570100, China.
*Corresponding author. Email: wsl@hainnu.edu.cn

REFERENCES AND NOTES

1. S. Kanwal, "Study of ground displacements, erosion and sea level rise along Karachi coast, Pakistan," thesis (The Hong Kong Polytechnic University, 2022).
2. H. Aeman, H. Shu, S. Abbas, H. Aisha, M. Usman, *Sci. Total Environ.* **880**, 163356 (2023).
3. A. C. Ferreira, R. Borges, L. D. de Lacerda, *Sustainability* **14**, 1263 (2022).
4. H. Gilani *et al.*, *Estuar. Coast. Shelf Sci.* **249**, 107128 (2021).
5. H. Brammer, R. Brinkman, in *Developments in Soil Science, Vol. 20: Soils on A Warmer Earth*, H. W. Scharpenseel, M. Schomaker, A. Ayoub, Eds. (Elsevier, 1990), pp. 145–156.
6. M. J. Kennish, *Environ. Conserv.* **29**, 78 (2022).
7. D. J. Curnick *et al.*, *Science* **363**, 239 (2019).
8. Z. Yang, P. Li, W. Feng, Y. Chen, *Adv. Function. Mat.* 10.1002/adfm.202305866 (2023).
9. R. Sunkur, K. Kantamaneni, C. Bokhoree, S. Ravan, *J. Sea Res.* **196**, 102449 (2023).
10. D. A. Friess *et al.*, *One Earth* **5**, 456 (2022).
11. A. Alamgir, M. A. Khan, N. Fatima, S. U. Fatima, *Environ. Monit. Assess.* **195**, 1308 (2023).
12. F. Wang *et al.*, *Natl. Sci. Rev.* **8**, nwaa296 (2021).

10.1126/science.adl3073

Light pollution driven by climate change

Global light pollution caused by artificial lighting at night increased by 7 to 10% per year between 2011 and 2022 (*1*). In addition to increasing the risk of chronic diseases in humans (*2*), light pollution can attract or repel insects and birds, with the potential to reshape their movement and alter habitat selection and distribution (*3*), and can harm marine species and ecosystems (*4*). Climate change–driven heat waves (*5*) have led to increased human activities at night, when temperatures drop, which has created more light pollution in vulnerable areas. Light pollution policies should limit additional lighting, despite increasing heat waves.

Recent heat waves in the Middle East have led Dubai's authorities to open beaches at night, when the temperatures are lower. To facilitate the recreational activities of nighttime visitors, floodlights illuminate the beach (*6*). This additional lighting could adversely affect coastal species, communities, and ecosystems. For example, light pollution has already been shown to affect nocturnally migratory birds, nocturnal pollinators, and turtles that inhabit these ecosystems (*7–9*).

About 600 million people in the Middle East and North Africa are likely to experience increasingly frequent extreme heatwaves in the second half of 21st century (*5*), with temperatures reaching or exceeding 56°C for weeks (*5*). Such temperatures might result in the populations in these regions shifting to more nocturnal activities. The increased artificial lighting required to support this lifestyle could substantially affect local species.

Establishing a robust network of light pollution monitoring in the Middle East is vital (*10*). Regulation to limit additional lights should be put in place immediately (*2*, *11*). To mitigate ecological effects, a precautionary approach should be applied to the size, color, type, and intensity of permissible lights. Large-area lighting in coastal areas and beaches should be prohibited and replaced with smaller devices, such as flashlights and occasional streetlights, during nighttime recreation. Filtered yellow-green and amber light-emitting diode (LED) lamps, which may have weaker effects on wildlife compared with high pressure sodium lamps, could be given preference, whereas blue-rich lighting, which may have the strongest effects, could be barred (*12*). Because light pollution effects are species specific (*8*), steps should be taken to minimize the intensity of all types of light. Finally, to provide local species with enhanced protection from artificial light, additional shrubs and trees should be planted in these areas to act as shelter.

Evgenios Agathokleous
Research Center for Global Changes and Ecosystem Carbon Sequestration & Mitigation, School of Applied Meteorology, Nanjing University of Information Science and Technology, Nanjing 210044, Jiangsu, China.
Email: evgenios@nuist.edu.cn

REFERENCES AND NOTES

1. D. C. M. Kyba, Y. Ö. Altintas, C. E. Walker, M. Newhouse, *Science* **379**, 265 (2023).
2. K. M. Zielinska-Dabkowska, E. S. Schernhammer, J. P. Hanifin, G. C. Brainard, *Science* **380**, 1130 (2023).
3. A. K. Jägerbrand, K. Spoelstra, *Science* **380**, 1125 (2023).
4. A. Wilkinson, *Science* 10.1126/science.aac6904 (2015).
5. G. Zittis *et al.*, *npj Clim. Atmos. Sci.* **4**, 20 (2021).
6. "In sweltering Dubai, beaches are open all night," *CNN* Youtube Channel (2023); https://www.youtube.com/watch?v=vl_L66XKcwQ&ab_channel=CNN.
7. S. A. Cabrera-Cruz, J. A. Smolinsky, J. J. Buler, *Sci. Rep.* **8**, 3261 (2018).
8. E. S. Briolat, K. J. Gaston, J. Bennie, E. Rosenfeld, J. Troscianko, *Nat. Comm.* **12**, 4163 (2021).
9. E. Silva *et al.*, *J. Photochem. Photobiol. B* **173**, 240 (2017).
10. M. Kocifaj, S. Wallner, J. C. Barentine, *Science* **380**, 1121 (2023).
11. M. Morgan-Taylor, *Science* **380**, 1118 (2023).
12. T. Longcore *et al.*, *J. Exp. Zool.* **329**, 511 (2018).

10.1126/science.adk7733

PRIZE ESSAY

GRAND PRIZE WINNER

Michael A. Skinnider

Michael Skinnider received his undergraduate degree from McMaster University and a MD/PhD from the University of British Columbia, during which time he was also a visiting PhD student at the École Polytechnique Fédérale de Lausanne. He then started his laboratory in the Lewis-Sigler Institute for Integrative Genomics and the Ludwig Institute for Cancer Research at Princeton University. His research focuses on the application of machine learning to problems in biology, chemistry, and medicine. **www.science.org/doi/10.1126/science.adk8626**

NOMIS & Science YOUNG EXPLORER AWARD

New designer drugs present unknown risks to users and treatment challenges to first responders.

SOCIAL SCIENCES

Hallucinating hallucinogens

Fighting the designer drug epidemic with generative AI

By **Michael A. Skinnider**[1,2]

Early on the morning of 12 July 2016, paramedics were called to the scene of a mass casualty event in Brooklyn (*1*). Bystanders had reported that multiple individuals were displaying unusual behavior that they described as "zombie-like." Investigators quickly established that all the affected individuals had smoked a packet of "herbal products" marketed as "AK-47 24 Karat Gold." But this was not a typical drug overdose: Most recreational drugs, after all, do not cause users to stagger mechanically through the street, moaning and staring blankly. What "herbal product" could cause such alarming symptoms?

Forensic scientists suspected the packet contained a completely new "designer drug" that had just emerged on the illicit market. Over the following weeks, a team of forensic chemists worked feverishly to establish the identity of the drug: a potent synthetic cannabinoid named AMB-FUBINACA. This compound had originally been developed by Pfizer but was later abandoned and never tested in humans. In vitro data suggested that AMB-FUBINACA was 85 times more potent than Δ^9-tetrahydrocannabinol, the principal psychoactive constituent of cannabis—a level of potency that could account for its zombie-like effects (*2*).

AMB-FUBINACA is just one of a wave of designer drugs that have reshaped the illicit drug market over the past two decades. Conventional drugs of abuse such as cocaine or methamphetamine still dominate the market, but enterprising chemists have realized that with slight chemical modifications to these drugs, they can create new derivatives that are completely legal (*3*).

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. [2]Ludwig Institute for Cancer Research, Princeton University, Princeton, NJ, USA. Email: skinnider@princeton.edu

But because most of these derivatives have never been tested in humans, they can have unpredictable and even fatal effects. This situation presents enormous challenges to first responders, who must treat completely new toxidromes, as well as public health officials, who must determine how to respond to outbreaks caused by uncharacterized molecules (*4*).

If these drugs could be identified faster, emergency physicians could provide better treatment to intoxicated patients, and public health officials could better guide responses to outbreaks. But identifying a new designer drug is no small feat. The most commonly used analytical method—mass spectrometry—requires a pure sample of the drug of interest for comparison. Yet pure samples are rarely available for drugs that have just emerged on the illicit market, and without pure samples, the identification of a new drug can take months.

As a MD/PhD student, I saw firsthand how patients could present with devastating symptoms of designer drug intoxications, but emergency physicians had few options to treat them. I wondered whether artificial intelligence (AI) could help. Specifically, I asked whether AI could automatically elucidate the chemical structures of new designer drugs from mass spectrometry data. Scientifically, this was a tall order; structure elucidation of unknown molecules with mass spectrometry is widely seen as an impossible task. But advances in generative AI made me optimistic. A new class of machine-learning models, called language models, had just begun to attract chemists' attention (*5*). Originally designed to parse human language, these models could instead be trained on textual representations of chemical structures called SMILES (simplified molecular input line entry system) (*6*). Once trained, language models could generate entirely new chemical structures that resembled the molecules in the training set. This phenomenon has been referred to as "hallucination" because the generated SMILES do not correspond to known chemical structures but instead reflect what the neural network views as plausible extensions of the training set (*7*). I thought that these models might overcome one of the primary challenges in structure elucidation: namely, generating unknown molecules to match the mass spectrometry data.

To test this hypothesis, I first had to overcome a major technical challenge. Previous language models had been trained on millions of chemical structures. But a crowdsourced effort to assemble the structures of all known designer drugs had come up with just 1753 molecules (*8*). It was unknown whether language models could even learn from so little data.

Unexpectedly, my colleagues and I found that with a slight tweak, language models could learn from as few as 1000 molecules (*9*). The key insight was to change how the molecules were represented—specifically, that the same chemical structure can be represented with many different SMILES (*10*). We found that this redundancy could provide a powerful mechanism for data augmentation: By enumerating many different SMILES for each molecule, the model could be tricked into thinking it was learning from a larger dataset of molecules.

Using this data augmentation trick, we successfully trained a language model on the structures of all 1753 known designer drugs and showed that this model could generate entirely new molecules whose structural properties were indistinguishable from those of known drugs (*11*). Intriguingly, these molecules were not all equally likely to be generated by the model. In a sample of 1 billion generated molecules, most chemical structures appeared just once or twice, but a tiny minority were generated tens of thousands of times.

This observation raised an exciting possibility: Could the model be learning to statistically anticipate the designer drugs that were most likely to emerge on the illicit market next? With the help of the Danish national forensic laboratory, we tested this hypothesis on a designer drug that had just been discovered: a new derivative of the street drug PCP (phencyclidine) named deoxymethoxetamine. I showed that given only the mass of this drug—one of the cheapest and easiest properties to measure experimentally—the language model could correctly elucidate deoxymethoxetamine's entire chemical structure (*11*).

As a final step, we integrated the language model's predictions with mass spectrometric data. We showed that combining these two sources of information allowed the model to elucidate the chemical structures of 40 new designer drugs (*11*). These results suggest that language models could dramatically accelerate the pace at which emerging drugs are identified.

I have now applied this technology to tens of thousands of patient samples and used it to discover several new designer drugs, such as a new analog of fentanyl that emerged last year. Currently, I am working with the British Columbia Centre for Disease Control to implement this AI technology in routine clinical practice to automatically discover new drugs as soon as they are introduced into the population. Ultimately, my dream is that first responders, emergency physicians, and public health officials will all be able to take advantage of generative AI to make more informed decisions when treating patients and managing outbreaks. ■

REFERENCES AND NOTES

1. A. J. Adams *et al.*, *N. Engl. J. Med.* **376**, 235 (2017).
2. S. D. Banister *et al.*, *ACS Chem. Neurosci.* **7**, 1241 (2016).
3. E. Underwood, *Science* **347**, 469 (2015).
4. E. Gebel Berg, *ACS Cent. Sci.* **2**, 363 (2016).
5. B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **361**, 360 (2018).
6. D. Weininger, *J. Chem. Inf. Model.* **28**, 31 (1988).
7. I. Anishchenko *et al.*, *Nature* **600**, 547 (2021).
8. M. Mardal *et al.*, *J. Anal. Toxicol.* **43**, 520 (2019).
9. M. A. Skinnider, R. G. Stacey, D. S. Wishart, L. J. Foster, *Nat. Mach. Intell.* **3**, 759 (2021).
10. E. J. Bjerrum, arXiv:1703.07076 [cs.LG] (2017).
11. M. A. Skinnider *et al.*, *Nat. Mach. Intell.* **3**, 973 (2021).

10.1126/science.adk8626

FINALIST

Isabella S. Bower

Isabella Bower received her undergraduate and postgraduate degrees in architecture from Deakin University and the University of Melbourne. After completing a PhD in psychology from Deakin University, she was awarded a prestigious postdoctoral research fellowship from the University of South Australia in 2023. Her research explores whether we can improve brain functioning and mental health through built environment design. **www.science.org/doi/10.1126/science.adk8629**

FINALIST

George Goshua

George Goshua received his MD from the University of Kentucky and completed his clinical residency and fellowship training in internal medicine and hematology-oncology at Yale University under the guidance of T. P. Duffy and A. I. Lee. He later completed an MSc focused on risk and decision science at the Harvard T.H. Chan School of Public Health under the supervision of A. Pandya. He opened his laboratory in the Department of Internal Medicine at the Yale School of Medicine in November 2022. His research applies health decision science methodology to identify equity-informed, cost-effective diagnostic and therapeutic treatment strategies for individuals living with hematologic and oncologic diseases.**www.science.org/doi/10.1126/science.adk8633**

PHOTOS: (TOP TO BOTTOM) COURTESY OF ISABELLA BOWER; JOEL CALLAWAY

PRIZE ESSAY

FINALIST

Isabella S. Bower

Isabella Bower received her undergraduate and postgraduate degrees in architecture from Deakin University and the University of Melbourne. After completing a PhD in psychology from Deakin University, she was awarded a prestigious postdoctoral research fellowship from the University of South Australia in 2023. Her research explores whether we can improve brain functioning and mental health through built environment design. **www.science.org/doi/10.1126/science.adk8629**

NOMIS & Science YOUNG EXPLORER AWARD

SOCIAL SCIENCES

How do our environments affect our minds?

Neuroscience reveals how building design shapes our behavioral, brain, and body responses

By **Isabella S. Bower**

I approach my research in an interdisciplinary manner because that is how my mind works. When I was growing up, I looked for patterns in my environment to help me make sense of my surroundings. When I was 2 years old, my parents took me on a car trip through Tasmania, an island state southeast of mainland Australia. I had not started speaking; yet, out of nowhere, I began singing "E-I-E-I-O" from the nursery rhyme Old MacDonald Had a Farm. This refrain turned into a reoccurring performance at random intervals, which caused my parents some confusion. Then they detected a pattern. The melody chimed from the backseat each time that we passed signage for a McDonald's restaurant. Often, the melody alerted them before they had noticed the sign along the road. A few years later, my kindergarten teachers reported that I would spend hours creating my own environments with wooden building blocks. Unlike other children, who would flit between activities, I fixated on incrementally building my own world and filtered out all other distractions around me.

What does my childhood have to do with my research? These stories paint a picture of how I see, process, and work with information about my environment to create new understanding. By the time I reached the end of high school, I wanted to combine my interests in design and science in my future education but could not see how. Two decades later, I found a way to combine science and design through my doctoral and now my postdoctoral research (see the photos).

Much like my early interest in synthesizing environmental information through an unexpected method (i.e., musical rhythm) and an incremental approach to understanding complexity (i.e., building blocks), I take the same approach to solving problems today. Although we spend most of our time inside buildings—whether in our homes, offices, classrooms, or health settings—we know unexpectedly little about the impact of our surroundings on our mental health and behavior. Perhaps this lack of knowledge is because we recognize just how complex the environment is, or maybe we simply have not taken the time to pause and reflect.

My research aims to understand the impact that design characteristics of the built environment have on cognitive processes, such as attention, perception, and emotional regulation. To achieve this, I combine psychology (behavior), neurobiology (brain), and physiology (body) data and methods with engineering (virtual reality) and insight from architecture (design practices) to analyze the impact of the buildings that we occupy on cognitive, emotional, and behavioral outcomes. Notably, this incremental and methodologically rigorous approach enables me to separate confounding variables, which previous work in this emergent field had not considered (*1*). These confounders include the role of factors such as physiological comfort through internal and external factors on the body (e.g., temperature, humidity, airflow, lighting, and acoustics). I have taken a careful approach to the role of past experiences and context too—delicately balancing context neutrality with realism. This approach recognizes that to properly understand the impact of our surroundings, we first need to understand what influences the occupants experiencing them.

My doctoral research tested the impact of the scale and color of an enclosed indoor room on self-report of emotion, physiological measures (e.g., heart rate, respiration rate, and sweat response), and brain waves (through electroencephalography, which measures electrical activity from brain cell communication). During this study, I recorded indoor environmental quality and used Cave Automatic Virtual Environment, an immersive form of virtual reality, to control a participant's surroundings. To understand the impact of room scale, I used the Australian Standards national building code to model a door, which provided a control from which I could then create enlarged and

UniSA Justice and Society Academic Unit, University of South Australia, Adelaide, SA, Australia.
Email: isabella.bower@unisa.edu.au

I have gone from drawing emotions (top; 2 years old in 1993) to using neuroscience to measure them in the present day (bottom; 32 years old in 2023).

PHOTOS: (TOP TO BOTTOM) VICKI BOWER; CATHERINE LEO

reduced conditions. To compare the difference between scale and another design characteristic, I added a blue color condition to compare with white. I chose blue because it is not often associated with behaviors or actions, which makes it less complex to distinguish whether the color or the associated behavior produces an impact.

The findings revealed that enlarged scale of the built environment affects brain activity linked to attention and cognitive performance (*2*), blue color modulates brain and body activity linked to emotional processing (*3*), and both color and scale change how different parts of the brain "talk" to one another (*4*). If scale affects brain activity linked to certain cognitive processes, does this result in behavioral changes? To test this hypothesis, I led a follow-up study using the same experimental design with a new cohort but with an important development—this time, participants performed a working memory and an emotion recognition task. I discovered that the reduced-scale room resulted in significantly faster response times for correct answers compared with the control room in the working memory task. The results suggest that the scale of interior built environments affects our ability to perform neurocognitive tasks.

Currently, our environments are built with an evidence base to support only the physical needs of occupants. My research aims to raise awareness and establish an evidence base for building design with the intended psychological health and performance needs in mind. Once unraveled, this emergent area of research holds immense potential for society, where it can be translated to inform built environment practice. This includes understanding how the built environment might affect different individuals—an important consideration to ensure that the buildings we design are truly inclusive and supportive for everyone.

Because people perform multiple cognitive tasks during their days, there may not be one ideal environment; rather, it may be important to have a range of accessible environments from which to choose based on what is most useful for each task. Notably, this research suggests that we may be able to optimize cognitive function for specific tasks (e.g., sustained concentration or pattern detection) and needs (e.g., reducing anxiety and aggravation or providing emotional comfort to others) through design, which could lead to enhanced productivity, well-being, and satisfaction.

I am grateful for those who have recognized and encouraged my divergent approach and perspective to understanding the world around me. Interdisciplinary research is challenging, but it is also highly rewarding. I am excited to inspire the next generation of researchers to build on this emergent field of research, advancing our knowledge of how our environments affect our minds. ■

REFERENCES AND NOTES

1. I. Bower, R. Tucker, P. G. Enticott, *J. Environ. Psychol.* **66**, 101344 (2019).
2. I. S. Bower *et al.*, *eNeuro* **9**, ENEURO.0104-22.2022 (2022).
3. I. S. Bower *et al.*, *Psychophysiology* **59**, e14121 (2022).
4. I. S. Bower, A. T. Hill, P. G. Enticott, *Hum. Brain Mapp.* **44**, 447 (2023).

ACKNOWLEDGMENTS

I.S.B. is currently supported by a postdoctoral research fellowship from the University of South Australia. During her PhD, she was supported by a postgraduate research scholarship from Deakin University, a fellowship from the Academy of Neuroscience for Architecture, and a Creativity and Innovation grant from Creative Futures Pty Ltd.

10.1126/science.adk8629

PRIZE ESSAY

FINALIST

George Goshua

George Goshua received his MD from the University of Kentucky and completed his clinical residency and fellowship training in internal medicine and hematology-oncology at Yale University under the guidance of T. P. Duffy and A. I. Lee. He later completed an MSc focused on risk and decision science at the Harvard T.H. Chan School of Public Health under the supervision of A. Pandya. He opened his laboratory in the Department of Internal Medicine at the Yale School of Medicine in November 2022. His research applies health decision science methodology to identify equity-informed, cost-effective diagnostic and therapeutic treatment strategies for individuals living with hematologic and oncologic diseases.**www.science.org/doi/10.1126/science.adk8633**

NOMIS & Science YOUNG EXPLORER AWARD

SOCIAL SCIENCES

Napkin math can change the world

Mathematical models hold the key to equitable patient care

By **George Goshua**

In 2017, as an internal medicine resident, I found myself taking care of a hospitalized patient with immune thrombotic thrombocytopenic purpura (TTP), a rare relapsing and remitting autoimmune-mediated hematological disease that did not have US Food and Drug Administration (FDA)–approved therapeutic options at the time. For more than a decade, leading academic medical centers across the globe had used the anti-CD20 monoclonal antibody rituximab off-label to help decrease the time-variant probability of relapse. However, my request for rituximab to treat immune TTP was denied, and the reason was cost—each dose was more than $7000. What this financial decision failed to account for is that even one repeat hospitalization is far more expensive, and this patient would surely return to our hospital for care in the case of relapse. The solution to this quantitative problem required nothing more than a back-of-the-envelope approach, incorporating the salient variables. Our "napkin math" was later published and used in support of inpatient formulary approvals of rituximab for immune TTP in the United States (*1*, *2*). Today, I use state-of-the-art methods in quantitative decision science in my laboratory, yet napkin math remains the initial step that determines the further pursuit of any idea and/or cross-discipline, multi-institutional collaboration.

Despite investment in health care spending that far outpaces that of any other country, the United States continues to rank at the bottom on overall performance, and across several subdomains of health care systems, as delineated by The Commonwealth Fund (*3*). I started my laboratory in November 2022 to work at the interface of clinical hematology-oncology, rare diseases, and health economics. Our work is focused on trying to understand how we can make better use of the financial embarrassment of riches in our health care system. In all my studies, which purposefully span the spectrum of understudied rare diseases, I strive to assemble teams with recognized clinical content and policy expertise to ensure construct validity. This approach has produced a collaboration network of more than a dozen institutions across the United States and Canada, which continues to expand.

Globally, there are ~200,000 individuals affected by the rare autoimmune cytopenia known as immune thrombocytopenia, which has both clinical bleeding and thrombotic manifestations. In 2021, I led a four-institution multidisciplinary team to fill a gap that we identified in the American Society of Hematology guidelines regarding this condition. In an even-if analysis, we showed that despite the past two decades of steady annual decrease in splenectomy use as a treatment modality from 30% to less than 5%—due to perceived tripartite risk of infection, thrombosis, and perioperative mortality and in lieu of chronic, expensive medication—the quality-adjusted life-expectancy value of this treatment for patients is, at minimum, on par with chronic medication therapy (*4*). In 2023, a talented first-year medical student in my laboratory was awarded an American Society of Hematology grant to expand this work on the value of splenectomy in warm autoimmune hemolytic anemia, which is another autoimmune cytopenia.

In a recent multi-institutional collaboration, my laboratory worked with colleagues at the Harvard T.H. Chan School of Public Health to quantitatively incorporate health equity into the treatment decision of gene therapy versus standard-of-care for patients with sickle cell disease (*5*–*7*), which is another rare disease, a hemoglobinopathy, that wreaks havoc on all organ systems. The course of the disease starts at an early age, due to an underlying genetic mutation, which leads to a loss in the structural integrity of red blood cell membranes and worsens with physiological and pathological stressors. We structured our recent article on the disease's treatment as a playbook by which other clinician-scientists can use distributional cost-effectiveness analysis methodology to account for health disparities salient to the health care treatment landscape of specific rare (as well as common) diseases (*5*). We are, at present, expanding on this work—within and beyond sickle cell

Yale University School of Medicine, New Haven, CT, USA.
Email: george.goshua@yale.edu

disease—with a multiyear effort to produce a Health Equity Atlas, which is a resource that will equip stakeholders with transparent numerical equity benchmarks as they make future coverage decisions.

My laboratory has also used quantitative decision science to safeguard against inefficacious new therapeutic modalities by showing the health economic consequences of clinically shallow primary end point–based FDA approval of such therapeutics, which are particularly rife in the hematology-oncology arena. A prime example is the push for the use of caplacizumab, an anti–von Willebrand factor agent, for unselected patients with immune TTP. This drug was approved by the FDA in 2019 on the basis of a laboratory primary outcome of improved time to platelet normalization. However, the drug lacked improvement in 8 of 10 outcomes that were deemed most important by the disease-specific expert panel that authored immune TTP guidelines in 2020 (*8*). The remaining two outcomes were disease exacerbation and relapse, ranked ninth and fifth in priority, and were shown to be better and worse, respectively, with caplacizumab use. I led a multi-institutional, cross-disciplinary analysis to demonstrate the cost ineffectiveness of this agent, a finding that was in stark incongruence with that of the overwhelming key opinion leader in support of this therapeutic agent (*9*). Our work generated two international debates in 2022 in London and New Orleans and informed dozens of formulary decisions across the United States (*10*–*12*). By 2023, real-world outcome data across all international reports collectively showed an unacceptable major bleeding risk with this medication (*2*, *13*). Even more-recent data show a need for longer therapy duration in one-third of patients who had caplacizumab added to their standard of care versus those who had the standard of care alone (*14*).

Leading a laboratory at the interface of quantitative decision science and clinical hematology-oncology is a risky career investment for me, which is rooted in the belief that we can do much more with the world's richest, but ultimately underperforming, health care system. I believe that data-driven improvement across key performance indicators needs to be directed from within by health care providers. My hope is that my work as a first-generation clinician-scientist inspires the next generation of clinician-scientists to discover the field of quantitative decision science and that its power lies in the explicit quantification of uncertainty and the ability to isolate the most-salient variables that affect clinical practice change. And, yes, it all starts with napkin math. ■

REFERENCES AND NOTES

1. G. Goshua, A. Gokhale, J. E. Hendrickson, C. Tormey, A. I. Lee, *Blood Adv.* **4**, 539 (2020).
2. A. Butt, C. Allen, A. Purcell, S. Ito, G. Goshua, *J. Clin. Med.* **12**, 4887 (2023).
3. E. C. Schneider *et al*., "Mirror, mirror 2021: Reflecting poorly" (The Commonwealth Fund, 2021).
4. G. Goshua *et al*., *Am. J. Hematol.* **98**, 122 (2023).
5. G. Goshua *et al*., *Ann. Intern. Med.* **176**, 779 (2023).
6. N. Lapid, "Sickle cell cure will be cost-effective if health disparities considered, study finds," *Reuters*, 15 December 2022.
7. R. Cookson, *Ann. Intern. Med.* **176**, 866 (2023).
8. X. L. Zheng *et al*., *J. Thromb. Haemost.* **18**, 2496 (2020).
9. G. Goshua *et al*., *Blood* **137**, 969 (2021).
10. A. Kumar, A.-R. Jeong, "A CAPtivating debate," *ASH News Daily* (2022).
11. G. Goshua, "Debate: Most/all patients with acquired (immune) thrombotic thrombocytopenia purpura receive caplacizumab. Education Spotlight," 64th Annual American Society of Hematology meeting, New Orleans, LA, 10 to 13 December 2022.
12. G. Goshua, V. Prasad, A. I. Lee, P. K. Bendapudi, *Lancet Haematol.* **8**, e315 (2021).
13. G. Goshua, P. K. Bendapudi, *Hematology Am. Soc. Hematol. Educ. Program* **2022**, 491 (2022).
14. N. Prasannan *et al*., *Blood* **141**, 2206 (2023).

10.1126/science.adk8633

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

HERBIVORY

Unequal impacts

Herbivory is a major selection pressure on plants, which have evolved many different physical and chemical adaptations to prevent animals from eating their tissues. However, herbivory pressure can be highly variable, even between plants in the same population. The Herbivory Variability Network consortium used standardized surveys to compare herbivory variability within populations at 790 sites across five continents. They found that the weak increase in mean herbivory at lower latitudes was accompanied by lower variation between individuals. Smaller plant species had higher herbivory variability, which also showed a phylogenetic signal. These findings highlight how variation in species interactions can influence eco-evolutionary outcomes. —BEL

Science, adh8830, this issue p. 679

The impact of herbivores, such as these birch sawfly larvae, varies largely within plant populations.

PHOTO: PAUL HOBSON/MINDEN PICTURES

MACHINE LEARNING

More confidence in machine learning

Over the past decade, there has been rapid progress in the development of large-scale machine learning (ML) systems that provide predictions related to various scientific phenomena. Unfortunately, the standard statistical approaches used to calculate confidence intervals and *P* values from gold standard data lose their statistical validity for ML-derived data. Angelopoulos *et al*. introduced "prediction-powered inference," a standardized protocol for constructing valid confidence intervals and *P* values that enables the power and scale of ML systems to be used as predictors while ensuring responsible and reliable scientific inferences. The method has been demonstrated on a broad range of real datasets and offers a promising statistical approach for using ML to derive scientific conclusions responsibly. —YS

Science, adi6000, this issue p. 669

ATMOSPHERIC SCIENCE

Aerosols' effect on low cloud coverage

Aerosols cool the climate indirectly by increasing low cloud brightness and coverage, partly offsetting anthropogenic greenhouse warming. However, the magnitude of the effect is uncertain, especially with cloud coverage. Yuan *et al.* used global ship tracks and remote sensing data to quantify aerosol-induced changes to marine low cloud properties through contrasting ship-polluted clouds to nearby background clouds, finding a strong observed aerosol effect on low cloud coverage. They conclude that low cloud coverage adjustments to aerosols adds between 52 and 300% additional forcing to that from low cloud brightness adjustments over the ocean. —JCHC

Sci. Adv. (2023) 10.1126/sciadv.adh7716

PHYSICS

Harnessing polyatomic molecules

Cold molecules hold much promise as a platform for investigating fundamental problems in physics such as matter-antimatter asymmetry through precise measurements of the electron's electric dipole moment (eEDM). Researchers have used diatomic molecules for this purpose. Polyatomic molecules have even more favorable properties but are more difficult to control. Anderegg *et al*. demonstrate coherent control of individual quantum states in the polyatomic molecule calcium monohydroxide. Their findings enable the measurement of eEDM in this and related systems, which is expected to lead to improved precision. —JS

Science, adg8155, this issue p. 665

PHOTONICS

Integrated mode-locked laser

Mode-locked lasers are an enabling technology in the ultrafast sciences, providing a platform to generate extremely short pulses of coherent light

and precisely spaced frequency combs of light. These lasers are typically bulky, with components sitting on an optical bench. Guo *et al*. shrunk a mode-locked laser down to the size of an optical chip. Combing a III-V gain medium with a lithium-niobate phase modulator, they demonstrated the operation of a mode-locked laser with good performance metrics. The results show promise for developing photonic chip–based frequency combs for precision measurements and spectroscopy. —ISO

Science, adj5438, this issue p. 708

ANTHROPOLOGY

Range expansion of *Homo erectus*

In 1981, the mandible of an infant hominin was discovered in the highlands of Ethiopia. Since then, its affinities have been debated, limiting our understanding of its importance. Using synchrotron x-ray imaging, Mussi *et al*. identified it as *Homo erectus*. This fossil is dated to 2 million years ago, and it is associated with both Oldowan and Early Archeulean tools, confirming that *H. erectus* used both types. This finding also reveals the use of high-elevation environments in this species, providing insight about the movement of hominins out of Africa. —SNV

Science, add9115, this issue p. 713

PHAGE ECOLOGY

Microcosm at multiscale

Real communities are characterized by complex interaction networks among species. The interaction is described as "nested" if there are specialist species interacting with subsets of species and "modular" if species interact within a group but not between different groups. A mixed pattern is known as a "multiscale" network. Borin *et al*. examined if and how a multiscale interaction network can develop through rapid evolution in a simple community with one type of bacteria and one type of phage. The authors were able to recapitulate the range of host interactions using receptor knock-out experiments. This means that phage-host interactions in a laboratory setting are sufficient to form complex ecological patterns and could be a valuable model system for informing phage therapy. —CA

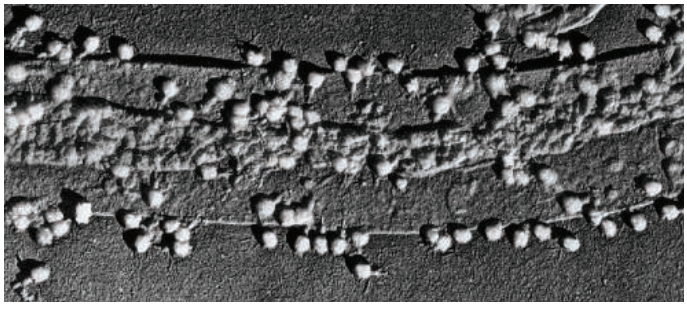
Science, adi5536, this issue p. 674

IMMUNE REGULATION

T cells protect muscle during exercise

Regulatory T cells (T_{regs}) support repair of injured muscle, but whether they participate in the response of healthy muscle to exercise training remains unclear. Using acute and chronic models of exercise in mice, Langston *et al.* found that T_{regs} suppress exercise-induced skeletal muscle inflammation that is counterproductive for performance enhancement. T_{regs} were required for gains in exercise capacity and promoted muscle metabolic reprogramming by protecting mitochondria from interferon-γ–driven damage. These results identify T_{regs} as a key regulatory element that is activated in response to exercise and needed to support performance-enhancing muscle adaptations. —CO

Sci. Immunol. (2023)
10.1126/sciimmunoladi5377



Electron microscopy image showing bacteriophages infecting an *Escherichia coli* cell

IN OTHER JOURNALS

Edited by **Caroline Ash** *and* **Jesse Smith**

STEM EDUCATION

Reimagining graduate handbooks

Graduate STEM education programs contain the same elements: research, coursework, and a candidacy examination. However, recent shifts in higher education research suggest that it is time to reconsider the goals of graduate programs. Donkor and Harshman approached this reframing through the lens of the graduate handbook. Using a backward design paired with document and thematic analysis, graduate handbooks from 60 chemistry departments were investigated for evidence of explicit alignment among the goals, objectives, and activities of the program. The results showed that most handbooks do not explicitly state learning goals, indicating missed opportunities for helping students value programmatic elements, identifying students' weaknesses, and helping faculty align their assessment tools with the learning goals. Including explicit learning goals and outcomes in graduate handbooks could improve student experiences and outcomes. —MMc

J. Chem. Educ. (2023)
10.1021/acs.jchemed.3c00062

VIROLOGY

Getting out of the nucleus

During infection, herpes simplex virus (HSV-1) assembles new viruses inside the host cell nucleus. To reach the nuclear envelope for viral exit, newly formed viral capsids must cross dense peripheral host heterochromatin. Lewis *et al.* observed increases in the levels of the epigenetic markers H3K27me3 and macroH2A1 on host cell chromatin during HSV-1 infection. Loss of these markers correlated with lower viral titers, reduced capsid mobility, and trapping of viral capsids in the nucleus. On infection, broad host genomic regions showed redistribution of macroH2A1 and H3K27me3 and decreased host transcription. Thus, virus-stimulated changes to host cell heterochromatin

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

HYPOTHESIS

How do molecular motors loop DNA?

Structural maintenance of chromosomes complexes (SMCs) are molecular motors that fold DNA into the three-dimensional structures that are important for many genomic processes. However, the mechanism by which these motors act is unclear. In a Perspective, Dekker *et al.* draw on recent experimental findings to present a new model for the extrusion of loops of DNA by SMCs. Their "reel-and-seal" model integrates central features of previous hypotheses but explains how SMCs can pass large obstacles on the DNA. —SAL

Science, adi8308, this issue p. 646

BIOELECTRONICS

Removing limits on powering devices

Bioelectronic devices that can sense or manipulate biological signals, such as pacemakers that can detect and regulate irregular cardiac cycles, can dramatically improve the health and lifestyle of the patients who use them. However, these devices are often limited by the storage capacity of an onboard battery or are tethered to wires that can cause infection. Nair *et al*. reviewed advances in developing alternate methods to generate, transmit, and store electrical charge that enable wireless power transfer and energy harvesting safely through and within the human body. In addition to removing power limitations, these advances often also enable communication from a device or possibly even between devices. —MSL

Science, abn4732, this issue p. 660

CORONAVIRUS

Inspiring inhibitors

The COVID-19 pandemic prompted many groups to rethink how to perform and communicate science. Although their work has been freely available from the beginning, Boby *et al*. now formally report the results of the COVID Moonshot project, a fully open-science drug discovery campaign to identify, synthesize, and test inhibitors against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) main protease, a key antiviral target (see the Perspective by Shoichet and Craik). Starting with data from a fragment-based screen, candidate inhibitor designs were crowdsourced from volunteer submitters using a variety of design approaches. An experienced team, aided by computational tools, evaluated proposals and designed synthetic routes. Noncovalent, nonpeptidomimetic inhibitors were identified and characterized functionally and structurally. Iterative medicinal chemistry and community input yielded a lead with promising bioavailability, safety, and antiviral activity. —MAF

Science, abo7201, this issue p. 663; see also adk5868, p. 649

METABOLISM

Absorb it with flowers

Intestinal absorption of dietary cholesterol has a major influence on blood cholesterol concentration, which plays an important role in cardiovascular and metabolic health. It was already known that this process starts with the Niemann-Pick C1 Like 1 protein, but now Ferrari *et al*. have identified subsequent steps involved in cholesterol uptake from the intestine. In particular, the authors determined that proteins called Aster-B and Aster-C are required for cholesterol to get from the exterior of intestinal cells to their endoplasmic reticulum, where it undergoes further processing. This pathway is targetable with a small-molecule inhibitor of Aster proteins, suggesting a potential pharmacological approach for lowering the intake of cholesterol. —YN

Science, adf0966, this issue p. 664

MAGNETISM

Vibrational route to large magnetic fields

The manipulation and control of the optoelectronic properties of a material finds application across a range of fields. However, doing so by applying electric or magnetic fields can be slow and not always practical. Luo *et al*. have shown that chiral phonons driven by ultrafast pulses of circularly polarized terahertz radiation can induce magnetic fields on the order of one tesla in the rare earth trihalide cerium fluoride (see the Perspective by Kaindl). Such control of spin-phonon coupling provides a route to on-demand ultrafast, large magnetic fields on an atomic scale that would be useful for both fundamental materials science and the development of energy-efficient spintronic devices. —ISO

Science, adi9601, this issue p. 698; see also adl3521, p. 642

WILDFIRES

Losing a home

Houses built near wildland vegetation are at greater risk of burning than those farther from the wildland-urban interface, a growing problem as housing developments expand and the climate becomes warmer. Radeloff *et al*. examined how the risk to houses within wildfire perimeters has changed in the US since 1990, finding that the number of homes within the wildland-urban interface and climate change are the most important factors controlling how many homes burn (see the Policy Forum by Boomhower). The number of homes destroyed by wildfires has doubled over the past 30 years, and most of them were in grasslands and shrublands, not near forests. —HJS

Science, ade9223, this issue p. 702; see also adk7118, p. 638

PLANT SCIENCE

Monitoring cell wall structure

Plant cells are surrounded by a polysaccharide cell wall that withstands internal cellular pressure but adapts and restructures as the cell expands. In growing pollen tubes, the proteins RALF4 and LRX8 are required to monitor integrity of the cell wall. Moussu *et al*. demonstrated that these proteins form a complex with pectic polysaccharides, the gel-like component of the cell wall (see the Perspective by Mohnen). In addition to the roles of RALF4 and LRX8 in signaling to the intracellular environment, the protein-polysaccharide complex performs a structural role in patterning the cell wall. This work contributes to our understanding of how signaling proteins can react to changes in the physical state of the extracellular structure. —MRS

Science, adi4720, this issue p. 719; see also adl1198, p. 648

MOLECULAR BIOLOGY

Male chromosomes' secret protector

Ultracompaction of sperm DNA in many species is often associated with genome-wide replacement of histones with protamines, but the actual role of this radical change in chromatin organization remains largely enigmatic. Looking at a multidecade-old *Drosophila* mutant, Dubruille *et al*. found that histones were massively retained in sperm without affecting its ability to fertilize (see the Perspective by Levine). However, at fertilization, male chromosomes were aberrantly recognized by maternal factors that control female meiotic progression, leading to a deleterious premature division and early loss of male chromosomes. This work highlights the role of sperm chromatin in protecting male chromosomes in the egg. —DJ

Science, adh0037, this issue p. 725; see also adl0365, p. 643

RADIATIVE COOLING

Cooling the hard way

Passive radiative cooling materials emit heat through the atmospheric window and into outer space, providing an attractive way to reduce temperatures in buildings. Zhao *et al*. created a passive cooling glass and Lin *et al*. developed a passive cooling ceramic, both of which are mechanically strong and relatively easy to scale (see the Perspective by Zhao and Tang). Unlike strategies that rely on polymers, these hard materials should be more robust to long-term weathering, which may make them far more useful for outdoor applications. —BG

Science, adi4725, adi2224, this issue p. 691, 684; see also adk9614, p. 644

BRAIN DEVELOPMENT

A window into cerebellum maturation

The human cerebellum undergoes a long maturation during early childhood and is especially susceptible to perturbations that contribute to the risk of developing neurodevelopmental disorders such as autism spectrum disorder. Ament *et al*. performed single-cell genomics on postmortem cerebellar samples from 1- to 5-year-old children who died from sudden accidental death or who experienced inflammation at the time of death. Gene expression changes were identified in Purkinje and Golgi inhibitory neurons in the samples from children with inflammation, including premature downregulation of the developmental genes required for cerebellar maturation. —OMS

Sci. Transl. Med. (2023) 10.1126/scitranslmed.ade1283

IMMUNOLOGY

Selectively GAGging chemokines

Chemokines induce the migration of immune cells by stimulating cognate chemokine receptors on the cell surface. Ridley *et al*. demonstrated how specificity is achieved in a system in which different chemokines can recruit the same immune cell types. The genes encoding the three ligands for the chemokine receptor CXCR3 were all expressed in the same context. Specificity was provided by the differential interactions of these chemokines with glycosaminoglycans, which regulated their localization and availability for signaling. —JFF

Sci. Signal. (2023) 10.1126/scisignal.adf2357

ICE SHEETS

Muddying the waters

Melting at the marine margins of the Greenland ice sheet can be caused by turbulent plumes of upwelling water formed by the discharge of meltwater from beneath the ice sheet, contributing substantially to sea-level rise. These plumes can contain significant amounts of sediment derived from erosion at the base of the ice sheet and transported by the basal meltwater flow. In a set of laboratory experiments, McConnochie and Cenedese showed that the melt rate induced by a particle-laden plume is greater than that caused by an equivalent plume without particles. Including this effect in models helps to reconcile predictions and observations of melting rates in these settings. —HJS

Geophys. Res. Lett. (2023) 10.1029/2023GL103736

Sediment in glacial discharge plumes increases the rate of melting at the margin of the Greenland ice sheet.

after Europeans arrived, resulting in 14 of 23 known species of parrots becoming extinct. —CNS

Proc. Natl. Acad. Sci. U.S.A. (2023) 10.1073/pnas.2301128120

PLANT SCIENCE

Stem cell viral exclusion

Many viruses that infect plants do not invade shoot stem cells, which prevents their spread into newly developing parts of the plant, including the germline. Incarbone *et al.* found that viruses are excluded from shoot apical meristems by the activities of a defense hormone, salicylic acid, which activates RNA interference proteins such as RNA-dependent RNA polymerases. The authors identified the tissue layers where the exclusion mechanism operates and found that these cells produce components that prevent the proliferation of many unrelated RNA viruses in stem cells. —MRS

Proc. Natl. Acad. Sci. U.S.A. (2023) 10.1073/pnas.2302069120

promote HSV-1 escape from the nucleus during productive infections. —SMH

J. Cell Biol. (2023) 10.1083/jcb.202304106

TASTE

Salmiakki takes getting used to

If you are from Scandinavia, chances are you learned to appreciate the unique taste of salty liquorice. This combination of a salty, bitter, and sour taste, "salmiakki," is produced by ammonium chloride (NH_4Cl). At high concentrations, ammonium NH_4^+ is toxic, and animals have evolved mechanisms to avoid potential poisons based on taste. However, how taste buds recognize NH_4^+ was unknown. Unexpectedly, Liang *et al.* have shown that a proton channel called OTOP1, which is expressed in sour taste buds, is the main NH_4Cl sensor, and an arginine residue in OTOP1 is responsible for NH_4Cl-mediated channel activation. These results suggest that salty liquorice must indeed be an acquired taste. —MMa

Some populations learn to relish the taste of licorice coated in ammonium chloride despite its toxic taste signal.

PHOTO: ETIENNE VOSS/ISTOCK.COM

Nat. Commun. (2023) 10.1038/s41467-023-41637-4

EXTINCTION

Polly wants a conservationist?

The arrival of humans in a new region has often been associated with biodiversity loss. However, disentangling factors leading to species loss is difficult because of how long ago they occurred. Oswald *et al.* used modern and ancient DNA, radiocarbon data, fossils, and archaeological data to estimate the timing and extent of loss of Caribbean parrot species. They found that parrots indeed displayed much greater genetic diversity historically and model phylogenies of extant and extinct species. Although some parrot species appear to have been lost after Indigenous settlement about 6000 to 1000 years ago, the most extensive loss occurred

BIOMATERIALS

Patterned vascularization using hydrogels

Although hydrogel scaffolds provide a water-rich environment for cell growth in reconstructive surgery, vascularization within the hydrogels tends to be slow and random. Ataie *et al.* developed a surgical micropuncture technique in conjunction with a granular hydrogel scaffold to guide and accelerate neovascularization through the early accumulation of macrophages and endothelial cells within the scaffolds. Perforations were created at the target blood vessel using an ultrafine microneedle onto which the granular scaffolds were deposited. The authors identified the optimum size for the gel particles, which correlated with the one that had the highest rate of macrophage infiltration. The induced microvasculature proved to be highly perfusable. —MSL

Small (2023) 10.1002/smll.202307928

REVIEW SUMMARY

BIOELECTRONICS

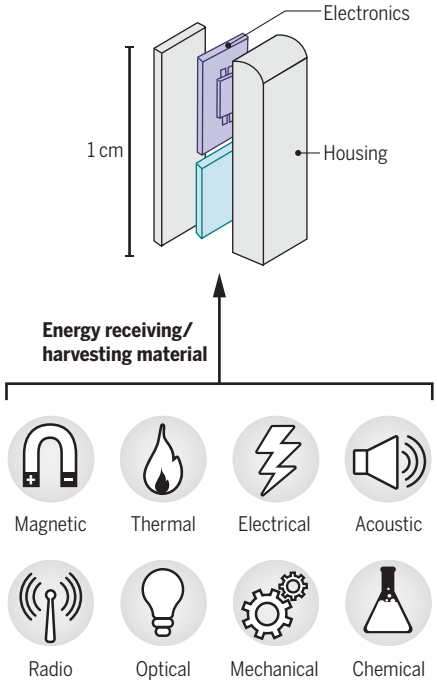
Miniature battery-free bioelectronics

Vishnu Nair, Ashley N. Dalrymple, Zhanghao Yu, Gaurav Balakrishnan, Christopher J. Bettinger, Douglas J. Weber, Kaiyuan Yang, Jacob T. Robinson*

BACKGROUND: Bioelectronics, electronic devices that sense or manipulate biological signals, can substantially improve medical outcomes by modulating the activity of excitable cells and tissues in the nervous system, cardiovascular system, and beyond. This specific stimulation and monitoring can reduce side effects and adapt in real time to the needs of each patient. However, despite the potential for bioelectronics to treat numerous conditions, only few implanted bioelectronic devices are currently the standard of care. These include cardiac pacemakers that use electrical impulses to synchronize the contraction of heart cells and continuous glucose monitors that generate an electronic alert when a user needs insulin.

Large battery packs, cumbersome tethers, and intricate packaging architectures complicate device design and present numerous possible failure modes, which limits the widespread deployment of therapeutic implantable bioelectronic devices. Wires can break and provide a pathway for infection. Furthermore, the need to route wires and leads limits the number and location of bioelectronic sensing and stimulating sites. Battery packs also create challenges for wearable bioelectronics such as heart rate monitors or pulse oximeters because the added size and weight of batteries limit the overall comfort, appearance, and ease of use.

Bioelectronic devices that operate without batteries could be miniaturized to millimeter-scale dimensions and easily implanted or comfortably worn. This approach would allow distributed networks of sensors and actuators to measure and manipulate physiological activity throughout the body to enable precise and adaptive bioelectronic therapies with minimal risks or interference with daily activities. To reach this level of miniaturization, bioelectronics must receive power from an external transmitter or harvest energy from the body itself. Unfortunately, traditional methods of transmitting wireless power over long distances rely on electromagnetic waves that are absorbed or reflected by the body. To overcome this challenge, recent innovations in wireless power transfer and energy-harvesting technologies have been leveraged to support the power requirements of bioelectronic devices. These include ultrasound, electromagnetic, magnetic, and optical methods to deliver power and data safely through the body, as well as energy-harvesting methods that extract chemical, kinetic, and thermal energy from the body itself. These emerging materials and technologies for data and power transfer are allowing bioelectronic devices to be miniaturized with tiny rechargeable batteries or, in the extreme case, made entirely battery free.



Battery-free bioelectronics. Miniature bioelectronics could support myriad therapeutics.

ADVANCES: We review progress in the materials and methods that are enabling wireless power transfer and energy harvesting safely through and within the human body. These methods must contend with the fact that many forms of energy are absorbed or reflected by bone and soft tissues, placing limits on the safety and efficiency of traditional power and data transmission technologies. To overcome these challenges, piezoelectric, photovoltaic, and magnetoelectric materials are being evaluated for use in nontraditional energy transfer techniques because they have unique advantages in the context of power transfer through the human body. Similarly, materials such as triboelectrics, magnetoelastics, and thermoelectrics are being explored to harvest energy from the human body. Furthermore, the emerging methods for power transfer through the human body are also being explored as a means for efficient data communication channels that will support wireless networks or miniature implantable bioelectronics.

OUTLOOK: The extreme miniaturization and long functional lifetime that are enabled by battery-free technologies point toward a future in which networks of tiny bioelectronics could be distributed throughout the body to accurately sense physiological states and apply therapies when and where they are needed. Reaching this vision will require additional scientific advances in manufacturing and packaging of the materials that support this wireless bioelectronic network. For example, to reach submillimeter scale, alternatives to traditional hermetic seals must be explored to protect these electronics from the harsh environment of the human body. In addition to typical material biocompatibility, implantable bioelectronics will have a longer operating lifespan if there is little to no foreign body tissue response and scar formation to degrade the recording or stimulation efficacy. Advances in nanomaterials, soft electronics, and coatings with growth factors, anti-inflammatory drugs, and peptides are underway, but ultimately these technologies need more in vivo testing throughout different regions of the body. This is especially critical if we are to propel bioelectronic devices into true distributed networks inside and nearby different organs and tissues throughout the body, each with their own movements, vasculature, immune cells, and repair mechanisms. Furthermore, distributed implanted bioelectronics will require specialized data and power transfer modalities that will need systems and standards for these emerging communication and energy transfer methods. For continued long-term success, power and data transfer modalities must be efficient, safe, secure, reliable, and able to be updated as new implants and methods arise. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jtrobinson@rice.edu
Cite this article as V. Nair *et al.*, *Science* **382**, eabn4732 (2023). DOI: 10.1126/science.abn4732

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn4732

REVIEW

BIOELECTRONICS

Miniature battery-free bioelectronics

Vishnu Nair[1], Ashley N. Dalrymple[2,3,4], Zhanghao Yu[5], Gaurav Balakrishnan[6], Christopher J. Bettinger[6], Douglas J. Weber[2,7,8], Kaiyuan Yang[5], Jacob T. Robinson[1,5]*

Miniature wireless bioelectronic implants that can operate for extended periods of time can transform how we treat disorders by acting rapidly on precise nerves and organs in a way that drugs cannot. To reach this goal, materials and methods are needed to wirelessly transfer energy through the body or harvest energy from the body itself. We review some of the capabilities of emerging energy transfer methods to identify the performance envelope for existing technology and discover where opportunities lie to improve how much—and how efficiently—we can deliver energy to the tiny bioelectronic implants that can support emerging medical technologies.

Electronic sensing and stimulation at multiple sites throughout the body enable bioelectronic therapies that go far beyond what is possible with pharmaceuticals. For example, by recording from sensory neurons in the dorsal root ganglia to control stimulation in peripheral nerves (*1*) or spinal cord, it is possible to restore walking in patients with severe spinal cord injuries (*2*). Electrical measurements of impedance from the colon can detect inflammation and be used to trigger the start of vagus nerve stimulation to reduce intestinal inflammation (*3*, *4*). Bidirectional bioelectronics can also be used to interact with external objects and the person's environment. For instance, implanted electromyography (EMG) electrodes can be used to control a myoelectric prosthesis (*5*, *6*). As the person interacts with and manipulates an object, sensors on the robotic arm can trigger stimulation through electrode arrays on or in the peripheral nerves to provide tactile feedback to the user (*5*, *6*). Alternatively, neural recordings from electrodes implanted into the motor cortex can be decoded to control a robotic limb, and sensory feedback can be provided by electrodes implanted in the somatosensory cortex (*7*).

For these types of medical advances to be deployed outside of controlled research settings, commercial bioelectronics must advance beyond the historical device architectures that are based on cardiac pacemakers developed in the 1950s (*8*, *9*). Indeed, the core components of the pacemaker architecture still exist in almost every commercial bioelectronic implant: a battery, a pulse generator circuit typically in the battery housing, and lead wires that connect the battery housing to the stimulation and/or recording site. Among these basic components, the battery remains the largest volume element in the system, so it prevents bioelectronic sensors and stimulators to be miniaturized to the size of injectables and distributed throughout the body, especially in sites where space is a premium. Recent improvements in battery technology and more power-efficient electronics have enabled more miniaturization, which provides hints of what is possible when the bioelectronic device can be small enough that it can be placed directly at the stimulation or recording site. For example, the Medtronic Micra is one of the first leadless pacemakers that integrates the battery, pulse generator, and electrodes into a single capsule that is small enough to be delivered to the heart through a catheter (*10*). This implant, roughly the size of a pen cap, is ~10 times smaller than modern pacemakers, which are roughly the size of an Oreo cookie (Fig. 1B). By combining the lead wires, battery, and electronics into this miniature package, the Micra avoids some of the primary failure modes related to lead wire migration, failure, or infection (*10*).

As promising as the Micra is, many applications require more power than cardiac pacing, making it difficult to create similarly sized miniature devices that are based on battery power. The power requirements of bioelectronic devices vary depending on the application. For example, cardiac pacing typically draws 100 to 700 μW of power (*11*, *12*), whereas tonic spinal cord stimulation of 2 to 8 V in amplitude, 60 to 100 Hz in frequency, and 100 to 500 μs in pulse width requires an estimated power of 24 μW to 3.2 mW (assuming stimulation electrodes with an impedance of 1 kΩ) (*13*). Deep-brain stimulation (DBS) stimulation typically requires 4 to 10 V in amplitude, 50 to 130 Hz in frequency, and 100 to 130 μs in pulse width, necessitating an estimated power of 80 μW to 6.5 mW (assuming stimulation electrodes with an impedance of 1 kΩ) (*14*). Thus, for stimulation applications that require microwatt-level powers, batteries must be significantly larger than those used for miniature cardiac pace makers such as the Micra.

Compared with stimulation, bioelectronic sensing can be supported with a power of 100 μW or less, and thus can be supported with smaller batteries, but extreme miniaturization will still require battery-free approaches. A modern pulse oximeter made using a flexible organic semiconductor consumes only 24 μW against an over-the-counter device consuming ~60 mW (*15*). Another sensing application that involves reading the neural biopotentials for seizure detection consumes only 140 μW of power (*16*). The total power consumption of the system must also include the supporting electronics, which often consume power in the range of 10 to 100 μW (*17*). Therefore, 200 μW is often a sufficient power budget for bioelectronics that only support sensing functions.

Battery-free technologies support extreme miniaturization

To achieve higher power levels without increasing the size of the implant, next-generation miniature bioelectronic devices must receive power from an external transmitter or from the body itself. We can quantify when battery-free solutions should be considered by calculating the volume of an implant for which batteries would be too small to support >24 hours of operation (V_{24h}) (Fig. 1C). In general, the amount of charge in an ideal battery is equal to the volume of the battery (V) multiplied by the volumetric charge storage capacity (k), thus we can write the battery lifetime (T) for a specific application drawing a current (I) as follows (*18*):

$$T = \frac{k}{I} \times V \quad (1)$$

Where k depends on the battery technology used. On the basis of Eq. 1, we can plot the battery lifetime versus battery volume for a given application (average current draw) and a given battery technology (Fig. 1C). We can then find the battery volume that would support a V_{24h} (Fig. 1C)

$$V_{24h} = 24 \times \frac{I}{k} \quad (2)$$

To reduce V_{24h} and enable smaller battery-powered devices, one must either increase the volumetric charge capacity of the battery or reduce the power draw on the battery. Figure 1C shows approximate V_{24h} for a hypothetical application drawing 8.3 mA current from a battery having a k = 0.05 (estimated from the data sheet of battery LITH 75 from Dantona

[1]Rice Neuroengineering Initiative, Rice University, Houston, TX, USA. [2]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [3]Department of Biomedical Engineering, University of Utah, Salt Lake City, UT, USA. [4]Department of Physical Medicine and Rehabilitation, University of Utah, Salt Lake City, UT, USA. [5]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. [6]Department of Materials Science & Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [7]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA. [8]Center for Neural Basis of Cognition, Pittsburgh, PA, USA.
*Corresponding author. Email: jtrobinson@rice.edu

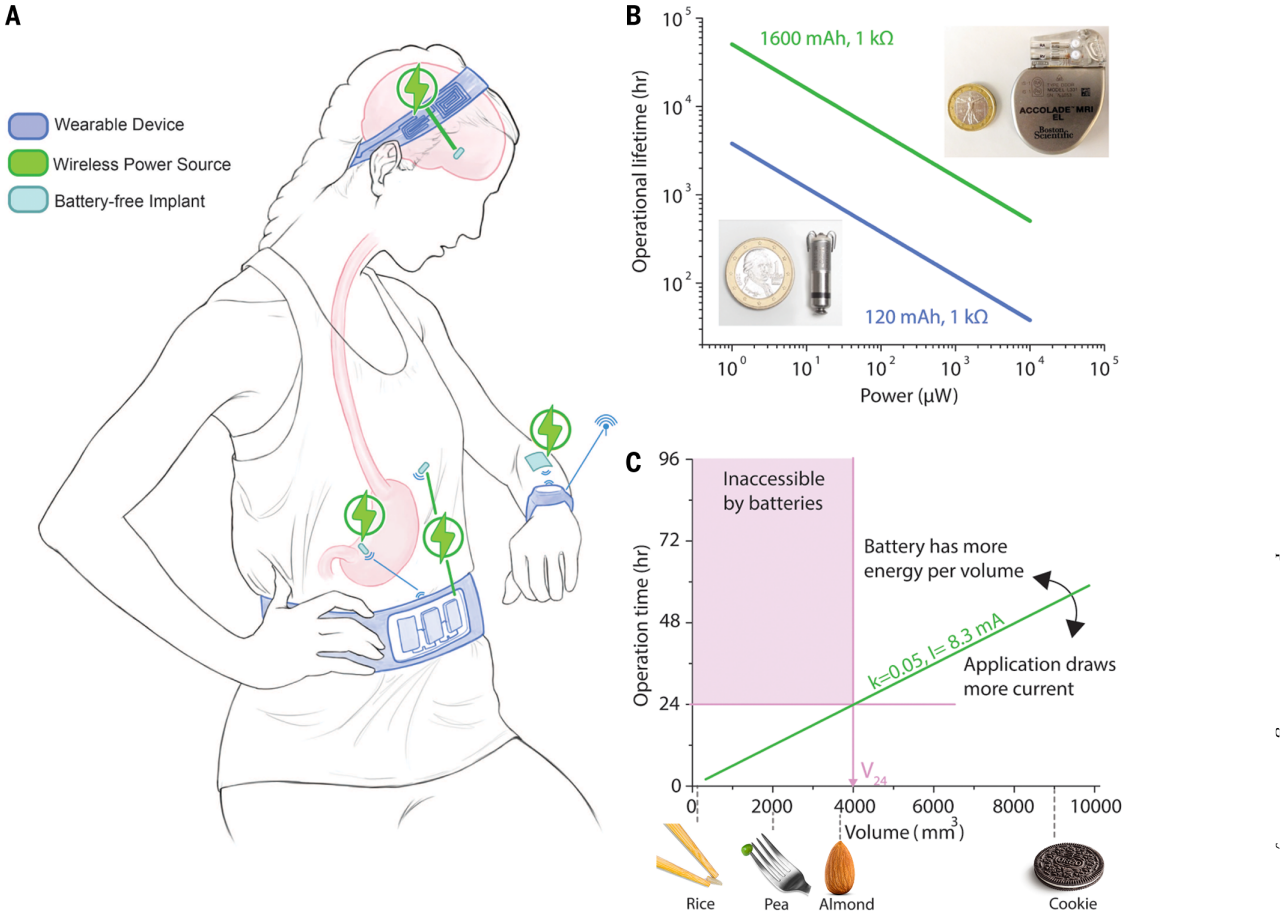


Fig. 1. Rationale for developing miniature battery-free bioelectronics. (**A**) Schematic of miniature battery-free bioelectronics that harvest energy from transmitters or the human body but also communicate wirelessly. [Illustration by Maayan Harel] (**B**) Operational lifetime as a function of power requirements for devices operating under 1 kΩ load. Example shown is a Boston Scientific Accolade pacemaker [photo courtesy of Vishnu Nair] with a 1600-mAh battery (green curve) and a Medtronic Micra pacemaker [photo reproduced with permission from (*116*)] with a 120-mAh battery (blue curve). Pacemaker photographs are shown with a 1 € coin (diameter = 23.25 mm) for size comparison. [Photos copyright Oxford Univ. Press, used with permission] (**C**) Operation time versus battery volume plot for a cylindrical commercial battery (lithium manganese oxide LITH-75 battery from Dantona Industries, k = 0.05 calculated from data sheet) drawing an 8.3-mA current. The pink horizontal line for an operation time of 24 hours meets the line defined by the power density and current draw at V_{24h}. Thus, for this example device, it is not possible to create an implant smaller than V_{24h} with that can last more than 24 hours between charges (pink shaded region). This zone is expected to be accessible only by devices powered wirelessly or through energy harvesting. Images of a rice grain (10 to 15 mm^3), pea (1 to 2 cm^3), almond (3 to 4 cm^3), and Oreo cookie (9 to 10 cm^3) are shown at the approximate locations along the x axis to indicate the sizes of the corresponding batteries ([Images are from iStock: almond, mrtekmekci; cookie, TheCrimsonMonkey; rice, dizainera; pea, nilsz]).

Industries). This analysis does not consider battery capacity fade over time, which would have increased V_{24h} further. Thus, to supply the energy that cannot be stored in a battery, we consider both energy harvesting from the human body and wireless power transfer (WPT) to devices on or in the human body. One challenge in making these comparisons, however, is that the figures of merit for batteries, energy harvesters, and WPT are all different because they scale differently as a function of size. Although batteries store more charge with increasing volume, energy harvesters create more power with increasing volume. We cannot compare these technologies without specifying the load they are driving. WPT technologies generally do not scale linearly with volume, so we must characterize the performance of batteries, energy-harvesting technologies, and WPT in terms of the figures of merit that best describe each technology. We can then compare these approaches by examining how technologies would perform if they were applied to the same bioelectronic application.

Our goal with this analysis is to understand the performance envelopes for existing technologies. This will enable us to explore current state-of-the-art technologies and future opportunities to develop materials and methods for energy transfer or harvesting that will expand the capabilities of battery-free bioelectronics.

Materials and methods for harvesting energy from the human body

Harvesting energy from thermal, mechanical, and chemical sources in the human body can support battery-free bioelectronics (Fig. 2), but there are several challenges that must be addressed before widespread adoption. These include a need for the components of the energy harvesting system to be biocompatible to prevent toxicity and inflammation. Further, the biological environment is aqueous and ionic, which creates the need to encapsulate devices to prevent material degradation or energy losses (*19*). Finally, with all technologies, there are manufacturing challenges associated with miniaturization.

Biomechanical energy harvesting

Triboelectric generators (TEGs) and piezoelectric generators (PEGs) offer two strategies to convert biomechanical work into electrical energy. TEGs harness the electrical field produced

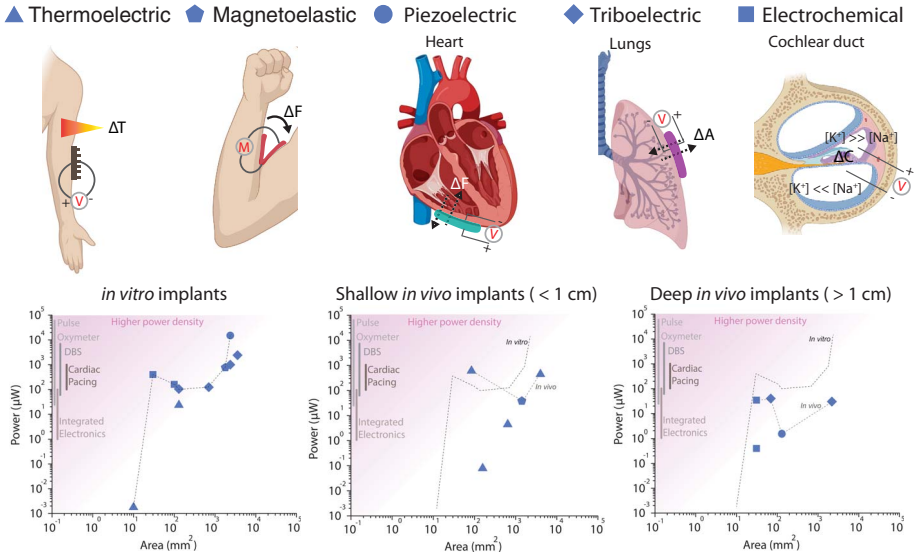


Fig. 2. Comparing the power densities of energy-harvesting technologies. Shown are schematic illustrations of thermoelectric (*43*, *46*, *117*–*120*), magnetoelastic (*47*), piezoelectric (*41*, *121*), triboelectric (*24*, *122*–*124*), and electrochemical (*35*, *39*) energy-harvesting technologies. The plots are divided into three groups: in vitro implants (*24*, *31*, *35*, *39*, *47*, *123*, *124*, *117*, *119*), which consist of devices with energy-harvesting capabilities that are measured by instrument-driven changes in the environment versus in vivo implants implanted in tissue or organisms. The in vivo implants are further divided into shallow implants in which the energy-harvesting element is implanted at a depth ≤1 cm and deep implants that are implanted at a depth of >1 cm. An envelope (dashed line) is constructed in each graph by joining the highest power-generating device for a given specific area. The in vitro envelope (black dashed line) overlayed with the two in vivo envelopes (gray dashed line) show that the performance of deeper implants is lower than those for their shallow and in vitro counterparts, suggesting room for improvement for these deeply implanted technologies. The *y* axis shows a plot of the range of powers needed for common bioelectronic applications such as pulse oximetry, DBS, cardiac pacing, and on-board integrated electronics. We expect this envelope to expand as new technologies are invented or existing technologies are improved to enable higher power densities. [Figure created with BioRender]

by contact electrification and charge separation. PEGs use materials that convert mechanical strain into an internal electric field. These generators can be used to harness voluntary mechanical work such as physical activity and involuntary mechanical work such as cardiac and circulatory activity. The typical power associated with natural limb motion of an 80-kg human walking at 1 Hz ranges from ~0.80 W in the elbow to ~34 W in the knee to ~1 W in the blood circulation (*20*). For powering bioelectronics, TEGs and PEGs must maximize conversion efficiency and power densities to achieve miniaturized form factors.

TEGs are created by manufacturing electron donors and acceptors into films, fibers, or other form factors that maximize charge separation and subsequent field generation. For deployment in bioelectronics, they typically use nontoxic materials that are amenable for implantable bioelectronics. Examples of electron acceptors include poly(dimethylsiloxane) (PDMS), poly (tetrafluoroethylene) (PTFE), and poly(vinylidene fluoride) (PVDF), whereas electron donors may include metal films (e.g., Ag, Au, Al, Cu) or polymers such as cellulose, silk, poly(ethyleneterephthalate) (PET), nylon, and poly(vinyl alcohol) (PVA) (*21*). Many of these materials have been used in implantable devices, which increases the likelihood for regulatory approval and eventual clinical adoption.

TEG prototypes typically adopt thin-film form factors with nominal footprints of 1 to 100 cm^2 that generate large peak voltages (>100 V) and low peak currents (10 to 100 μA) (*21*, *22*). There have been many advances in TEGs to increase areal and volumetric power densities and energy conversion efficiencies. These have included the synthesis and deployment of nanoscale generators in various architectures and modes of operation and optimization of materials selection (*22*). The resulting specific areal peak power generation can approach 10 mW/cm^2, but is typically ~0.01 mW/cm^2 (*23*, *24*). Areal peak power densities vary because of substantial variations in materials selection, fabrication approaches, and device structural design. These areal power densities during in vitro operation project to volumetric peak power densities of 1 to 10 mW/cm^3 (*23*, *24*). Further, TEGs can be optimized to operate at efficiencies as high as 55% (*25*).

The choice of materials with piezoelectric properties for PEGs is much smaller compared with those that exhibit triboelectricity. Prominent inorganic piezoelectric materials include exotic compounds such as lead zirconate titanate (PZT) and barium titanate (BT), which are challenging to implement in biomedical applications because of their intrinsic toxicity and brittle mechanical properties. Zinc oxide (ZnO) thin-film and aligned nanowire structures have been used as a biocompatible inorganic material with a strong piezoelectric response. For instance, a flexible ZnO nanowire array generated power densities as high as ~10 mW/cm^3 (*26*). The β phase of flexible, nontoxic PVDF fluoropolymers also exhibits a well-documented piezoelectric response (*27*, *28*). β-PVDF has been used to harvest mechanical energy and to power implantable bioelectronics since as early as 1984 (*29*). Centimeter-scale PVDF energy-harvesting patches (200 nm in thickness) generate a peak power of 40 nW in vivo (*30*). Thin-film PVDF devices are wrapped around blood vessels to measure arterial pressure, but the materials and device architecture could also be scaled as an energy-harvesting subsystem that can supply power to other bioelectronic devices. Additionally, hybrid piezoelectric and electret nanogenerators have been fabricated using various structures composed of biocompatible poly(L-lactide) (*31*). That study reported instantaneous maximum power densities of ~65 mW/m^2 for in vitro operation for application in wearable devices (*31*). More recently, a piezoionic response, or pressure-driven ionic flux, was engineered by patterning hydrogels with fixed charge densities, such as anionic polyacrylic acid within neutral polyacrylamide hydrogels to create voltage gradients on deformation. Power densities of up to 0.85 $\mu W/cm^3$ were generated and harnessed for fully hydrogel-based transducers capable of force sensing and direct peripheral nerve stimulation (*32*).

TEGs and PEGs of 1 cm^3 can theoretically meet the energy requirements of many implantable medical devices, including cardiac pacemakers (>100 μW), drug pumps (>0.1 mW), and neuromodulation devices such as cochlear implants (~1 mW) (*33*). Biomechanical generators that convert voluntary movements to power are limited by their dependence on physical activity and their sensitivity to humidity. As a result, the peak power generation varies substantially from average power production. Coupling these energy-harvesting approaches with energy storage systems or interfacing them with mechanically dynamic tissues such as the heart or blood vessels remains a promising power-harvesting strategy. To deploy biomechanical generators in chronic applications, there is also a critical challenge in designing functionally stable structures resistant to cyclic mechanical loading. The amount of cyclic fatigue, and therefore lifetime, depends on many factors such as the implant, site, the

device geometry, the materials chemistry, microstructure, strain rate, and maximum strain. A typical implant with a target chronic lifetime of 10 years is calculated to accommodate ~10^8 cycles at a deformation frequency of 1 Hz.

Electrochemical cells and other energy-harvesting strategies

Electrochemical cells that integrate biological components can be used for onsite power generation in bioelectronic devices. For instance, there have been attempts to harness an existing biologically maintained electrochemical gradient in the cochlea, known as the endocochlear potential, to power bioelectronic implants. An in vivo demonstration showed that the endocochlear potential of a guinea pig was able to provide ~6 nW of power with operating voltages near 55 mV (*34*). Although the power and voltages extracted are too low for most conventional bioelectronic applications, custom low-power integrated circuit design could increase the feasibility of using such power sources in specific cases. Another approach for biological batteries has been pursued, specifically to power transient ingestible bioelectronic devices with the gastrointestinal fluids serving as the electrolyte. Using magnesium or zinc anodes and copper cathodes, gastrointestinal batteries produced average power densities of 0.23 μW/mm^2 and were able to power ingestible central temperature sensors with wireless data transmission capabilities (*35*). A similar approach was used to design fiber-based injectable batteries with biofluid electrolytes in a variety of tissues such as the dermis, heart, and brain while producing power densities as high as ~80 mW/cm^3 (*36*). Practical limitations such as electrode degradation, large fluctuations in produced power, battery size and footprint, and device packaging still need to be addressed before the adoption of bioelectrochemical energy harvesting systems for bioelectronics (*37*).

Biofuel cells use biological motifs often immobilized on an anode that oxidize biomolecules and generate a flow of electrons to a corresponding cathode. These cells, including anodes with immobilized microbial and enzymatic species, have been explored as energy-harvesting devices for bioelectronic implants for >50 years (*38*). Enzyme-based biofuel cells such as glucose oxidase fuel cells are attractive as energy-harvesting systems for implantable devices because glucose is an abundant energy source. Advances in electrode materials and enzyme immobilization and stabilization have improved implantable glucose oxidase fuel cells that can operate stably for up to 1 year in vitro (*39*). The theoretical upper bound on volumetric current density of an enzyme-based fuel cell, assuming a specific activity of 1 U/mg and a mass density of 1 g/cm^3, is 1.6 A/cm^3 (*40*). However, a lower specific enzyme activity, suboptimal loading, and other losses reduce this projected in vitro current output to values between 1 μA/cm^3 and 1 mA/cm^3. Such a wide range of reported values for current density is attributed to the variability in initial enzyme activity, enzyme stability, and enzyme immobilization techniques combined with typical variances in electrode materials and device fabrication. The power density is also limited by the typically low open-circuit voltage (V_{oc}) of redox reactions occurring in aqueous environments and the thermodynamic potential window of water. These circumstances yield areal power densities of 0.1 to 1 mW/cm^2 for in vitro operation. Fuel cells that optimize enzyme packaging, maximize charge transfer at interfaces, and facilitate charge transport can achieve in vitro areal power densities of ~4 mW/cm^2 (*41*). However, practical limitations need to be addressed before the deployment of biofuel cells as energy-harvesting subsystems, including the efficient removal of reaction by-products, enzyme regeneration, and form factors to minimize fibrous capsule formation (*42*).

Thermoelectric generators (ThEGs) use specialized materials in which a thermal gradient promotes diffusion of charge carriers to create an electric potential. This need for a temperature gradient limits the use of ThEGs mainly to peripheral devices such as wearables or subdermal implants, locations where the temperature gradients between the body and the ambient environment are greatest (*43*). Thermoelectric materials are defined by their Seebeck coefficients S (specific magnitude of induced thermoelectric voltage), typically ~100 μV/K. A further figure of merit for thermoelectric materials is a dimensionless figure ZT ($ZT = \frac{S^2 \sigma T}{k}$), where σ is the electrical conductivity, T is the average temperature, and k is the thermal conductivity that should be maximized (*44*). Conventional thermoelectric materials are typically rigid, inorganic semiconductors such as bismuth-telluride, silicon and silicon-germanium alloys, and lead tellurides. Some of these materials, particularly bismuth-telluride and silicon, have been used as thick films, thin films, or nanomaterial building blocks for ThEGs in bioelectronic applications. For instance, a bismuth telluride ThEG harvested power at ~40 μW/cm^2 density to sustainably power a cardiac pacemaker (*45*). However, these materials have limited applications in vivo because of their toxicity and rigid mechanical properties. Conjugated polymers and ionic liquid-based ThEGs have shown promise for mechanically compliant ThEGs. More recent in vivo demonstrations of thermoelectric generators produce specific areal power densities that approach 1 mW/cm^2 (*46*). Thus, centimeter-scale energy-harvesting devices could meet the demands of many types of bioelectronic implants. Perhaps the most intriguing advantage of thermoelectric power generation for bioelectronics is that thermal gradients are always present in epidermal and subdermal applications. Although reasonable specific power densities are already achievable, materials with increased power factor of $S^2\sigma$ (*44*) could further increase the efficiency of thermoelectric energy-harvesting devices. It is essential to design systems that maximize the thermal gradient by selecting high ZT materials, using heat sinks, and optimizing device geometries and structures to prevent lateral heat loss. Further, the effectiveness of ThEGs varies by climate in which it is deployed. If efficient thermoelectric materials can be made that are mechanically flexible, nontoxic, and easily processable into miniaturized or thin-film form factors, they would be compelling candidates for generators for bioelectronic applications.

Magnetoelastic materials harness elastic deformation to produce local variations in magnetic field. A recent demonstration showed that a soft composite of magnetic iron-neodymium-boron microparticles in a silicone matrix produce a magneto-mechanical coupling factor of ~7×10^{-8} T /Pa, almost four times greater than rigid metallic alloys (*47*). Coupling composite materials with magnetic inductors allows soft magnetoelastic generators to supply ~20 W/m^2 from biomechanical energy for wearable applications. MEGs can achieve peak powers of 30 μW with ultrasonic stimulation.

Materials and methods for WPT technologies

For cases in which power harvested from the body cannot adequately power a miniature bioelectronic device, wireless power transmission can often achieve greater power densities. However, care must be taken to design the materials and methods so that this energy transfer is safe for the human body. The safety limitation results primarily from the fact that energy intended for a bioelectronic device can also be absorbed by the body, and this can lead to unsafe levels of tissue heating (*48*). Tissue heating–associated safety limits are set by specific absorption rates that depend on the conductivity of the tissue, density of the tissue, frequency of the electromagnetic wave, and root mean square of the electric field. Unfortunately, the high-frequency electromagnetic waves preferred for wirelessly powered electronics are relatively easily absorbed by the body (*49*). This has led to a renaissance in research to identify materials and methods to receive energy from nontraditional energy sources such as low-frequency magnetic fields, light, and ultrasound. These technologies have shown substantial improvements in the amount of power that can be received by miniature bioelectronic implants (Fig. 3) (*50*). Unlike batteries and energy harvesters, however, WPT depends on the size and orientation of the transmitter, the depth of the device in the tissue, and the size

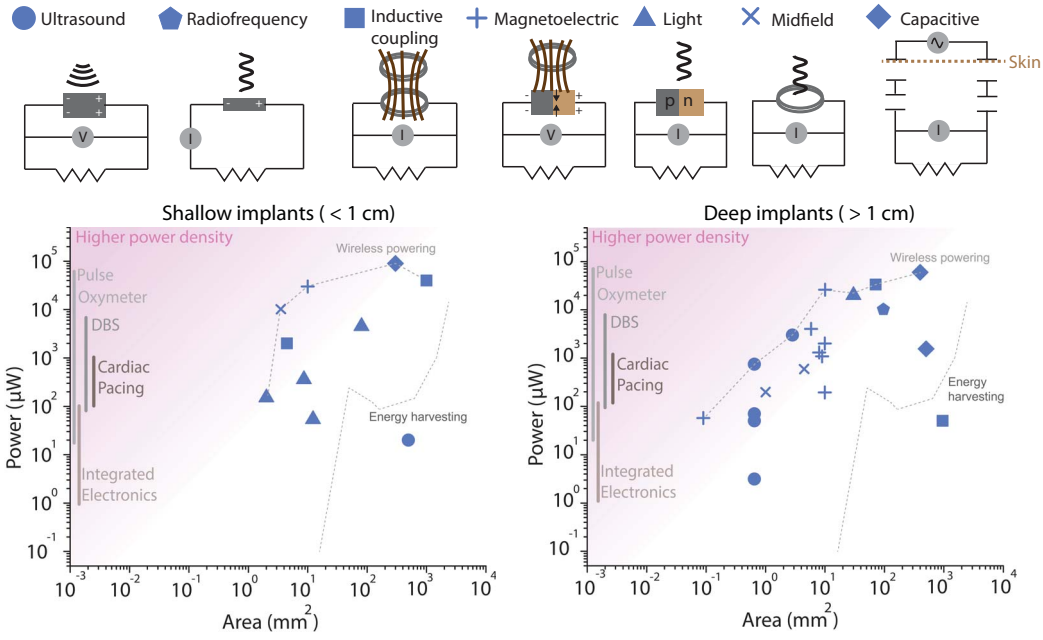


Fig. 3. Comparing the power densities of WPT technologies. Schematics of WPT technologies correspond to the plots below. Scatter plots show power versus receiver area for various WPT devices (*54*, *55*, *58*, *61*–*71*, *73*–*81*, *95*) classified into deep versus shallow implants, where a deep implant has stimulation electrodes or the receiver >1 cm below the surface of the skin. Shallow implants have the receiver element ≤1 cm below the surface of the skin. The light gray dashed line labeled "wireless powering" connects the highest power devices for each size. The dark gray dashed line labeled "energy harvesting" represents the in vitro envelope obtained from the corresponding analysis in Fig. 2. Here, we only consider demonstrations of wireless power within human safety limits. The *y* axis shows a plot of the range of powers needed for common bioelectronic applications such as pulse oximetry, DBS, cardiac pacing, and on-board integrated electronics. Technologies to the top left of these plots have the highest power densities. We expect this envelope to expand as new technologies are invented or existing technologies are improved to enable higher power densities.

of the receiver (*51*). Therefore, power density is only one of the many properties to consider when evaluating the best WPT technology for a particular application.

Near-field inductive coupling: The gold standard

Near-field inductive coupling (NIC) is by far the most common method for power delivery to biomedical implants and is the standard against which any new modality will be evaluated. This technique uses alternating magnetic fields generated by a coil of wire to induce an electric current in a receiver coil (*52*–*57*). The induced electric current can then be used to power the electronic elements. Efficient power transfer for inductive coupling depends on the total magnetic flux that is captured by the receiver coil (*52*–*57*). Thus, the efficiency in NIC with a millimeter-scaled receiver scales with the area of the receiver (A_R) and the separation between receiver and transmitter (S_{TR}) as follows: $A_R^{3/2} S_{TR}^{-3}$ (*58*). This leads to power transfer efficiencies <1% when the receiver is deeper than 1 cm under tissue (*58*). Inductive coupling is also sensitive to angular misalignment and axial separation of coils (*59*). The angular misalignment can be overcome to an extent by designing a transmitter coil with larger flux volume compared with the receiver coil and can be mitigated by using a ferromagnetic material in the center of the receiver coil (*52*–*57*). However, using ferromagnetic material in receivers can interfere with diagnostic methods such as computed tomography (CT) and magnetic resonance imaging (MRI) scans. Typical frequencies for inductive coupling are in the range of 1 to 13 MHz, which approaches the frequencies at which magnetic fields begin to be absorbed or reflected by the ionic fluids within the body (*60*). Although the safety of WPT is determined by the specific absorption rate, regulatory authorities such as the FCC allocate a spectrum of frequencies deemed safe for humans whenever the US Food and Drug Administration approves a new biomedical technology for clinical use. Currently, various bands within the range of 400 to 2400 MHz have been approved for medical radio communications, micropower and body area networks. Further the decay of the transmitter coil magnetic field along its axis limits the miniaturization of implanted devices to depths of 5 to 10 mm, for power requirements greater than 10 mW (*54*).

Emerging WPT techniques

There are several modalities of WPT that seek to improve upon NIC in terms of power density, depth, misalignment tolerance, and efficiency. These techniques have been reviewed extensively elsewhere (*51*). We compiled data from published work to determine how much power can be delivered as a function of the size of the receiver and compared that with the power levels needed for common bioelectronics applications. In Fig. 3, we plot the received power (in microwatts) versus receiver area (in square millimeters) for ultrasound (*61*–*63*), radiofrequency (*64*, *65*), magnetoelectric (*66*–*70*), light (*71*–*76*), midfield (*59*, *77*, *78*), and capacitive (*79*–*81*) power transfer with shallow implants (<1 cm in tissue) and deep implants (>1 cm in tissue). One centimeter is used as a reference to distinguish between deep versus shallow implant because it is approximately the depth from the scalp to the dura in an adult human or the thickness of skin across epidermis to the subcutaneous layer. Examining these plots revealed where new methods will be needed to achieve compact devices capable of high-power bioelectronic applications. For example, we found no deeply implanted battery-free bioelectronics implants smaller than ~1 mm^2 that could receive ~1 mW of power or more. Thus, improvements in the materials and methods are needed to make ultraminiature devices with the power needed for applications such as DBS and pulse oximetry. We also found numerous instances in which magnetoelectrics, light, ultrasound,

and radiofrequency demonstrated higher powers compared with similarly sized NIC-powered devices. These data support the trend that these emerging technologies can begin to replace NIC for miniature bioelectronic power delivery to enable more advanced functionality. For this analysis, we restricted the data to applications that fell within the safety limits for human exposure. It is also critical to acknowledge that there are many other metrics upon which one can compare WPT technologies. Alignment tolerance, power transfer efficiency, and comfort of the wearable transmitter could be more important factors when considering bioelectronic systems (*51*).

Wireless powering versus energy harvesting at a miniature scale

By plotting the received power versus the device size in Figs. 2 and 3, we could identify which bioelectronic applications can be enabled using emerging power technologies of different sizes. The dashed lines in Fig. 3 connect the highest-power devices at each size for WPT. A similar line plotted for in vivo energy-harvesting technologies shown in Fig. 2 reveals that state-of-the-art WPT achieves 100 to 1000 times more received power in miniature bioelectronic devices for both deep and shallow implants. This discrepancy is likely because energy-harvesting efficiency typical scales with device volume rather than device area. Therefore, without significant advances in energy-harvesting materials and methods, WPT will be the best solution for applications with high-power requirements. Nevertheless, when we look to the power levels required for applications such as cardiac pacing, DBS, and pulse oximetry plotted along the y axis (Fig. 2), we can see that millimeter-sized devices are beginning to meet the power requirements necessary to enable these types of applications with tiny implants.

Materials and methods for wireless communication and telemetry

For many applications, users and physicians must communicate with the bioelectronic devices both to reprogram their operation and to receive sensor data that can inform treatment. Wireless data transmission through or near the human body suffers many of the challenges related to wireless power transmission, namely absorption and reflection of radiating waves by the body. Fortunately, the same materials and methods used to overcome the wireless power transmission problem can also be used to send and receive data near or through the human body. This includes transmitting data to the devices (downlink) and receiving data from the device (uplink). Sharing the same method for power and data transmission is highly desirable for device miniaturization. However, considering various power and data rate targets in different applications, sharing the same method is not always feasible nor the optimal solution.

Downlink communication is typically used for infrequent device programming at low data rates. Therefore, almost all battery-free bioelectronic systems transmit downlink data by directly modulating the wireless power transmission signal. The simplest modulation scheme is on-off keying (OOK), which turns the power transmitter on and off to represent binary data. OOK downlink has been exploited with inductive coupling (*82*–*84*), ultrasound (*62*), and light (*85*) power transmissions. Amplitude-shift keying (ASK) modulation is commonly adopted in inductive coupling (*86*), radio (*87*), and magnetoelectrics (*67*, *68*). ASK encodes data on the amplitude of the power signal to avoid complete on/off switching cycles, thus increasing the data transmission rate and reducing energy costs. ASK, along with OOK, the simplest form of ASK, are both popular in downlink communications because they make the data demodulation easier, leading to lower power consumption and design overhead in the implants.

Uplink communication is more challenging because of the broad data rate and form factor targets for different sensing applications. The data rates needed to measure the dynamics of physiological signals such as neural activity and temperature, and biochemical concentrations can vary from around one to tens of thousands of kilobits per second. For example, measuring the spiking waveforms from one channel of neural activity typically requires sampling at 10 KHz, but this rate can be greatly compressed to 1 KHz if one were only interested in capturing principle components of the wave forms (*87*), down to 3.3 KHz to measure the timing of action potentials, or down to 300 to 1000 Hz to simply measure the spectral power in the spiking band (*88*–*91*). We summarized uplink schemes for wireless battery-free bioelectronics into three categories based on their relation to the wireless power transmission method. The major figures of merit for communication are the amount of energy (in picojoules) consumed per bit and the data rate (in megabytes per second). When multiplied together these constraints determine the amount of power consumed for data transmission. In Fig. 4, we plot these two figures of merit for several different communication strategies described below.

Passive backscattering

Passive backscattering is a low-energy wireless technology that exploits the physical process of reflecting waves coming from an external transmitter by modulating the input impedance of receivers at negligible energy cost. It is well suited for many bioelectronic applications because it exploits the highly asymmetrical energy budget between the battery-free device and the external transmitter. Passive

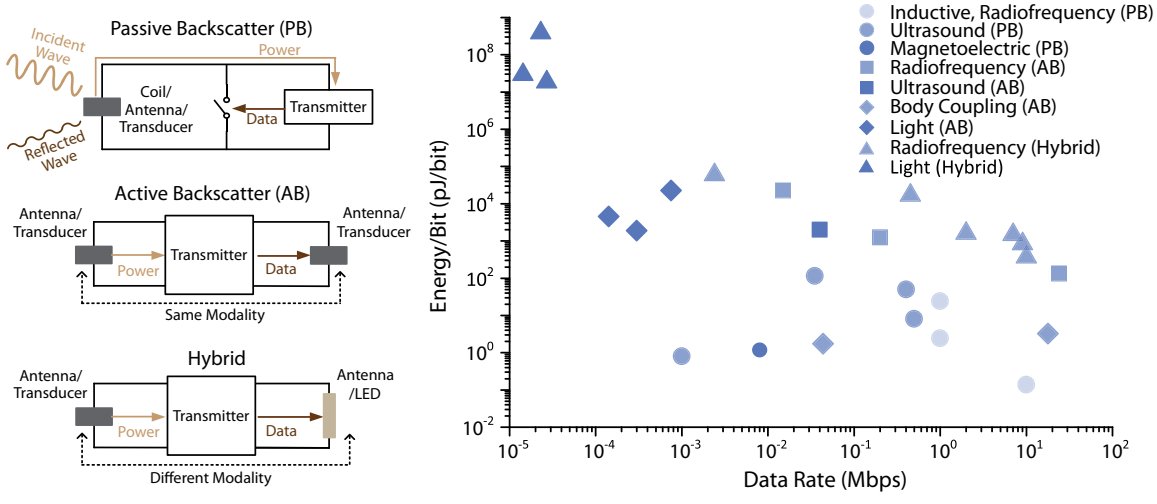


Fig. 4. Uplink communication mechanisms for bioimplants. Shown are recent methods demonstrated in vivo for uplink communication, along with their tradeoffs between energy and data rate. The communication schemes are categorized into passive backscattering (PB), active backscattering (AB), and hybrid mode depending on their relation to the wireless power technology.

backscattering of inductive coupling (*86*, *87*, *92*) and ultrasound waves (*61*, *62*, *93*) have been demonstrated in vivo for sensing applications. Recent work also demonstrated the feasibility of leveraging magnetoelectric converse effects for backscattering (*94*). Load shift keying is the most common modulation scheme to transmit digital binary data using different strengths of reflection. Specifically, changing the load impedance to an antenna, or simply shorting it in certain cases, will result in a change in the amplitude or frequency of the backscattered signal. Such load modulation can be easily implemented using a switch controlled by the uplink data bit, making it a straightforward and practical approach. The choice of load modulation method and receiver design must be optimized for a specific type of wireless signal. Beyond digital data uplink, directly modulating the reflected waves using amplified biosignals has shown promise for reducing the device size and power, leading to sub-cubic-millimeter neural recording devices (*62*, *95*). The limitations of such analog backscattering methods are lower signal-to-noise ratio and incompatibility with additional recording channels. As shown in Fig. 4, passive backscattering devices present the lowest energy consumption among all uplink technologies, with inductive backscattering having superior data rates because of its higher frequency compared with ultrasound.

Active backscattering

Active backscattering refers to the process of broadcasting uplink data using the same method as for wireless power transmission. Active backscattering is most common when wireless power transmission is performed using radio waves (*85*, *96*, *97*) or light (*71*, *85*). Radiofrequency electromagnetics provide a high bandwidth to transmit data at high speed, allowing multichannel biosignal recordings. Radiofrequency active backscattering is largely limited by low-power transmission due to tissue absorption. Further, most bioelectronic devices with optical power transfer adopt active backscattering because of the difficulty in reflecting light and the small footprint required by a data-transmitting LED. Additionally, ultrasound communication can be performed in the active backscattering approach by applying an alternating voltage to the energy-harvesting piezoelectric transducer to generate OOK-modulated ultrasound waves (*98*). In principle, active backscattering consumes more energy than its passive counterpart but has greater potential for higher signal-to-noise ratio and thus deeper implants. Finally, there is growing interest in tissue-coupled communication schemes that use biological tissue as a medium, which require only electrodes on the miniaturized implants. Conventional intrabody capacitive and galvanic communication schemes (*99*) have been modified and improved for communication to and from implants, such as biphasic quasistatic (*100*, *101*) and ionic (*102*) communication.

Hybrid communication and power transmission

The last category of bioelectronic devices uses different materials and methods for power and communication. Although integrating different mechanisms may increase the device size and assembly complexity, it offers the freedom to separately optimize power and communication links based on specific applications and naturally supports simultaneous power and data transmission. Radiofrequency transmitters are widely used in implantable bioelectronics that demand high bandwidth and are often combined with WPT through inductive coupling (*103*–*107*) and tissue energy harvesting (*34*). Commercial cardiac pacemakers and DBS systems use bluetooth low energy (BLE) for communication. This technology usually achieves a data rate of 1 Mbps, accompanied by a power consumption exceeding 1 mW. LED-based optical uplink is another popular uplink method in hybrid systems (*52*, *83*, *108*) because of its small footprint, simple design, and easy integration.

Outlook: Miniature bioelectronics embedded in wireless networks as the future

The wealth of energy-harvesting, WPT, and communication technologies provides a rich toolbox for creating next-generation battery-free bioelectronics that can be miniaturized and long-lasting in ways that battery-powered devices cannot. Energy-harvesting technologies provide major advantages because the human body can supply this power without the need for external transmitters. Energy harvesters demonstrated to date typically capture 10^{-3} to 10^{3} μW with physical footprints of 10^{1} to 10^{3} mm^2 (*109*). As shown in Fig. 3, the in vitro envelopes of energy-harvesting devices reveal power generation in the range of 10^{2} to 10^{4} μW over harvesting element areas of 10 to 10^{3} mm^2, corresponding to power densities of ~10 μW/mm^2 (*109*). However, deep implant in vivo devices can only generate powers in the range of 1 to 10 μW for similar sized devices with power densities in the range of 0.01 to 0.1 μW/mm^2 (*109*). Thus, the focus of future work should be toward designing devices to bridge this gap in power-harvesting capabilities. Although there are many applications in sensing that are supported at these power levels, stimulation at energies beyond cardiac pacing are typically orders of magnitude higher. In these cases, WPT may be suitable.

Current WPT technology envelopes (Fig. 4) reach power levels of ~0.1 to 100 mW in 0.1- to 1000-mm^2–sized devices, achieving power densities of 100 to 1000 μW/mm^2 (*109*). Thus, WPT power densities can generate real power densities 10 to 100 times those of ideal in vitro energy-harvesting devices. At these power levels, many therapeutic stimulation applications such as DBS and spinal cord stimulation become accessible. Combining stimulation with measurement will enable miniature and distributed networks for providing closed-loop therapies envisioned by bioelectronics (*110*). To facilitate real-time feedback control, implants should also be equipped with low-power uplink communications. Because transmitting data through the body typically expends between 0.1 and 52 pJ/bit at target data rates of 2.5 to 20.48 Mbps, we would expect instantaneous power consumption to range from 1 to 655 μW (*100*, *102*, *111*). The total power consumed by data transmission will depend on the quantity of data and the frequency of transmission required for a specific application, but instantaneous power places design constraints on closed-loop bioelectronic systems. Thus, for WPT technologies to support these types of close-loop systems, we would expect target power densities to be ~10^{3} μW/mm^2 and above. However, power densities cannot be used as the only measure to determine the practicality of a given WPT technology. Other factors factors such as transmitter size and weight, transmitter-receiver coupling, electromagnetic radiation safety, and implantability must also be considered when designing future bioelectronic systems.

In the near term, the most likely clinical applications for miniature battery-free implants will be less-invasive alternatives for existing neuromodulation and cardiac therapies. Miniature battery-free implants for cardiac pacing, pain, overactive bladder, sleep apnea, and mood disorders are all in commercial development (*112*–*115*). Other applications that do not require long-term implantation, such as post-operative monitoring, drug delivery, and cancer therapy, may be able to move more quickly through the regulatory approval process and reach the market in the next 5 years. Just as the battery-powered implanted pulse generators developed for cardiac pacing in the 1960s created a platform that enabled decades of innovation in bioelectronic medicine, miniature battery-free bioelectronics and networks of wireless implants may be the next great platform technology for the decade's innovation in bioelectronic medicine.

REFERENCES AND NOTES

1. B. J. Holinski, D. G. Everaert, V. K. Mushahwar, R. B. Stein, Real-time control of walking using recordings from dorsal root ganglia. *J. Neural Eng.* **10**, 056008 (2013). doi: 10.1088/1741-2560/10/5/056008; pmid: 23928579
2. T. M. Bruns, J. B. Wagenaar, M. J. Bauman, R. A. Gaunt, D. J. Weber, Real-time control of hind limb functional electrical stimulation using feedback from dorsal root ganglia recordings. *J. Neural Eng.* **10**, 026020 (2013). doi: 10.1088/1741-2560/10/2/026020; pmid: 23503062
3. S. C. Payne *et al.*, Anti-inflammatory effects of abdominal vagus nerve stimulation on experimental intestinal

inflammation. *Front. Neurosci.* **13**, 418 (2019). doi: 10.3389/fnins.2019.00418; pmid: 31133776

4. S. C. Payne *et al*., Transmural impedance detects graded changes of inflammation in experimental colitis. *R. Soc. Open Sci.* **7**, 191819 (2020). doi: 10.1098/rsos.191819; pmid: 32257338
5. J. A. George *et al*., Biomimetic sensory feedback through peripheral nerve stimulation improves dexterous use of a bionic hand. *Sci. Robot.* **4**, eaax2352 (2019). doi: 10.1126/scirobotics.aax2352; pmid: 33137773
6. D. W. Tan *et al*., A neural interface provides long-term stable natural touch perception. *Sci. Transl. Med.* **6**, 257ra138 (2014). doi: 10.1126/scitranslmed.3008669; pmid: 25298320
7. S. N. Flesher *et al*., A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science* **372**, 831–836 (2021). doi: 10.1126/science.abd0380; pmid: 34016775
8. S. K. Mulpuru, M. Madhavan, C. J. McLeod, Y.-M. Cha, P. A. Friedman, Cardiac pacemakers: Function, troubleshooting, and management. *J. Am. Coll. Cardiol.* **69**, 189–210 (2017). doi: 10.1016/j.jacc.2016.10.061; pmid: 28081829
9. L. D. Abrams, W. A. Hudson, R. Lightwood, A surgical approach to the management of heart-block using an inductive coupled artifical cardiac pacemaker. *Lancet* **1**, 1372–1374 (1960). doi: 10.1016/S0140-6736(60)91151-X; pmid: 13791559
10. N. Bhatia, M. El-Chami, Leadless pacemakers: A contemporary review. *J. Geriatr. Cardiol.* **15**, 249–253 (2018). pmid: 29915613
11. P. Abiri *et al*., Inductively powered wireless pacing via a miniature pacemaker and remote stimulation control system. *Sci. Rep.* **7**, 6180 (2017). doi: 10.1038/s41598-017-06493-5; pmid: 28733677
12. W. E. Hill, A. Murray, J. P. Bourke, L. Howell, R. G. Gold, Minimum energy for cardiac pacing. *Clin. Phys. Physiol. Meas.* **9**, 41–46 (1988). doi: 10.1088/0143-0815/9/1/003; pmid: 3359743
13. D. M. Moore, C. McCrory, Spinal cord stimulation. *BJA Educ.* **16**, 258–263 (2016). doi: 10.1093/bjaed/mkv072; pmid: 27458592
14. R. Ramasubbu, S. Lang, Z. H. T. Kiss, Dosing of electrical parameters in deep brain stimulation (DBS) for intractable depression: A review of clinical studies. *Front. Psychiatry* **9**, 302 (2018). doi: 10.3389/fpsyt.2018.00302; pmid: 30050474
15. H. Lee *et al*., Toward all-day wearable health monitoring: An ultralow-power, reflective organic pulse oximetry sensing patch. *Sci. Adv.* **4**, eaas9530 (2018). doi: 10.1126/sciadv.aas9530; pmid: 30430132
16. A. Bahr *et al*., Epileptic seizure detection on an ultra-low-power embedded RISC-V processor using a convolutional neural network. *Biosensors (Basel)* **11**, 203 (2021). doi: 10.3390/bios11070203; pmid: 34201480
17. P. M. Lee, Z. Xiong, J. Ho, Methods for powering bioelectronic microdevices. *Bioelectron. Med. (Lond.)* **1**, 201–217 (2018). doi: 10.2217/bem-2018-0005
18. N. E. Galushkin, N. N. Yazvinskaya, D. N. Galushkin, A critical review of using the Peukert equation and its generalizations for lithium-ion cells. *J. Electrochem. Soc.* **167**, 120516 (2020). doi: 10.1149/1945-7111/abad69
19. S.-W. Hwang *et al*., A physically transient form of silicon electronics. *Science* **337**, 1640–1644 (2012). doi: 10.1126/science.1226325; pmid: 23019646
20. R. Riemer, A. Shapiro, Biomechanical energy harvesting from human motion: Theory, state of the art, design guidelines, and future directions. *J. Neuroeng. Rehabil.* **8**, 22 (2011). doi: 10.1186/1743-0003-8-22; pmid: 21521509
21. Z. L. Wang, Triboelectric nanogenerators as new energy technology for self-powered systems and as active mechanical and chemical sensors. *ACS Nano* **7**, 9533–9557 (2013). doi: 10.1021/nn404614z; pmid: 24079963
22. B. Yang, Y. Xiong, K. Ma, S. Liu, X. Tao, Recent advances in wearable textile-based triboelectric generator systems for energy harvesting from human motion. *EcoMat* **2**, e12054 (2020). doi: 10.1002/eom2.12054
23. P. Bai *et al*., Integrated multilayered triboelectric nanogenerator for harvesting biomechanical energy from human motions. *ACS Nano* **7**, 3713–3719 (2013). doi: 10.1021/nn4007708; pmid: 23484470
24. Y. Sun *et al*., Flexible hybrid piezo/triboelectric energy harvester with high power density workable at elevated temperatures. *J. Mater. Chem. A Mater. Energy Sustain.* **8**, 12003–12012 (2020). doi: 10.1039/D0TA04612D
25. L. Lin *et al*., Robust triboelectric nanogenerator based on rolling electrification and electrostatic induction at an instantaneous energy conversion efficiency of ~55%. *ACS Nano* **9**, 922–930 (2015). doi: 10.1021/nn506673x; pmid: 25555045
26. S. Lee *et al*., Super-flexible nanogenerator for energy harvesting from gentle wind and as an active deformation sensor. *Adv. Funct. Mater.* **23**, 2445–2449 (2013). doi: 10.1002/adfm.201202867
27. S. Panda *et al*., Biocompatible CaTiO3-PVDF composite-based piezoelectric nanogenerator for exercise evaluation and energy harvesting. *Nano Energy* **102**, 107682 (2022). doi: 10.1016/j.nanoen.2022.107682
28. N. A. Shepelin *et al*., New developments in composites, copolymer technologies and processing techniques for flexible fluoropolymer piezoelectric generators for efficient energy harvesting. *Energy Environ. Sci.* **12**, 1143–1176 (2019). doi: 10.1039/C8EE03006E
29. E. Häsler, L. Stein, G. Harbauer, Implantable physiological power supply with PVDF film. *Ferroelectrics* **60**, 277–282 (1984). doi: 10.1080/00150198408017528
30. X. Cheng *et al*., Implantable and self-powered blood pressure monitoring based on a piezoelectric thinfilm: Simulated, *in vitro* and *in vivo* studies. *Nano Energy* **22**, 453–460 (2016). doi: 10.1016/j.nanoen.2016.02.037
31. S. Gong, B. Zhang, J. Zhang, Z. L. Wang, K. Ren, Biocompatible poly(lactic acid)-based hybrid piezoelectric and electret nanogenerator for electronic skin applications. *Adv. Funct. Mater.* **30**, 1908724 (2020). doi: 10.1002/adfm.201908724
32. Y. Dobashi *et al*., Piezoionic mechanoreceptors: Force-induced current generation in hydrogels. *Science* **376**, 502–507 (2022). doi: 10.1126/science.aaw1974; pmid: 35482868
33. H. A. Yiğit, H. Uluşan, M. B. Yuksel, S. Chamanian, B. Çiftci, A. Koyuncuoğlu, A. Muhtaroğlu, H. Külah, "A pulse-width modulated cochlear implant interface electronics with 513 μW power consumption," in *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Lausanne, Switzerland, 29–31 July 2019* (IEEE, 2019); doi: 10.1109/ISLPED.2019.8824838
34. P. P. Mercier, A. C. Lysaght, S. Bandyopadhyay, A. P. Chandrakasan, K. M. Stankovic, Energy extraction from the biologic battery in the inner ear. *Nat. Biotechnol.* **30**, 1240–1243 (2012). doi: 10.1038/nbt.2394; pmid: 23138225
35. P. Nadeau *et al*., Prolonged energy harvesting for ingestible devices. *Nat. Biomed. Eng.* **1**, 0022 (2017). doi: 10.1038/s41551-016-0022; pmid: 28458955
36. Y. Zhao *et al*., Injectable fiber batteries for all-region power supply *in vivo*. *J. Mater. Chem. A Mater. Energy Sustain.* **9**, 1463–1470 (2021). doi: 10.1039/D0TA10357H
37. T. J.-P. Ivase, B. B. Nyakuma, O. Oladokun, P. T. Abu, M. N. Hassan, Review of the principal mechanisms, prospects, and challenges of bioelectrochemical systems. *Environ. Prog. Sustain. Energy* **39**, 13298 (2020). doi: 10.1002/ep.13298
38. A. T. Yahiro, S. M. Lee, D. O. Kimble, Bioelectrochemistry: I. Enzyme utilizing bio-fuel cell studies. *Biochim. Biophys. Acta* **88**, 375–383 (1964). pmid: 14249845
39. B. Reuillard *et al*., One-year stability for a glucose/oxygen biofuel cell combined with pH reactivation of the laccase/carbon nanotube biocathode. *Bioelectrochemistry* **106** (Pt A), 73–76 (2015). doi: 10.1016/j.bioelechem.2015.04.009; pmid: 25934467
40. G. T. R. Palmore, G. M. Whitesides, in *Enzymatic Conversion of Biomass for Fuels Production*, M. E. Himmel, J. O. Baker, R. P. Overend, Eds. (American Chemical Society, 1994), vol. 14, pp. 271–290.
41. C. H. Kwon *et al*., High-power hybrid biofuel cells using layer-by-layer assembled glucose oxidase-coated metallic cotton fibers. *Nat. Commun.* **9**, 4479 (2018). doi: 10.1038/s41467-018-06994-5; pmid: 30367069
42. S. Cosnier, A. J. Gross, A. Le Goff, M. Holzinger, Recent advances on enzymatic glucose/oxygen and hydrogen/oxygen biofuel cells: Achievements and limitations. *J. Power Sources* **325**, 252–263 (2016). doi: 10.1016/j.jpowsour.2016.05.133
43. C.-G. Han *et al*., Giant thermopower of ionic gelatin near room temperature. *Science* **368**, 1091–1098 (2020). doi: 10.1126/science.aaz5045; pmid: 32354840
44. E. W. Zaia, M. P. Gordon, P. Yuan, J. J. Urban, Progress and perspective: Soft thermoelectric materials for wearable and internet-of-things applications. *Adv. Electron. Mater.* **5**, 1800823 (2019). doi: 10.1002/aelm.201800823
45. D. Bhatia, S. Bairagi, S. Goel, M. Jangra, Pacemakers charging using body energy. *J. Pharm. Bioallied Sci.* **2**, 51–54 (2010). doi: 10.4103/0975-7406.62713; pmid: 21814432
46. D. Rozgic, D. Markovic, A miniaturized 0.78-mW/cm^2 autonomous thermoelectric energy-harvesting platform for biomedical sensors. *IEEE Trans. Biomed. Circuits Syst.* **11**, 773–783 (2017). doi: 10.1109/TBCAS.2017.2684818; pmid: 28541912
47. Y. Zhou *et al*., Giant magnetoelastic effect in soft systems for bioelectronics. *Nat. Mater.* **20**, 1670–1676 (2021). doi: 10.1038/s41563-021-01093-1; pmid: 34594013
48. IEEE, "IEEE standard for safety levels with respect to human exposure to electric, magnetic, and electromagnetic fields, 0 Hz to 300 GHz," in *IEEE Standard C95.1-2019* (IEEE, 2019), *2019 Revision of IEEE Std C95.1-2005/Incorporates IEEE Std C95.1-2019/Cor 1-2019*); doi: 10.1109/IEEESTD.2019.8859679
49. F. T. Alrashdan *et al*., Wearable wireless power systems for 'ME-BIT' magnetoelectric-powered bio implants. *J. Neural Eng.* **18**, 045011 (2021). doi: 10.1088/1741-2552/ac1178; pmid: 34229314
50. R. Hinchet *et al*., Transcutaneous ultrasound energy harvesting using capacitive triboelectric technology. *Science* **365**, 491–494 (2019). doi: 10.1126/science.aan3997; pmid: 31371614
51. A. Singer, J. T. Robinson, Wireless power delivery techniques for miniature implantable bioelectronics. *Adv. Healthc. Mater.* **10**, e2100664 (2021). doi: 10.1002/adhm.202100664; pmid: 34114368
52. A. Burton *et al*., Wireless, battery-free subdermally implantable photometry systems for chronic recording of neural dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2835–2845 (2020). doi: 10.1073/pnas.1920073117; pmid: 31974306
53. G. E. Loeb, R. A. Peck, W. H. Moore, K. Hood, BION system for distributed neural prosthetic interfaces. *Med. Eng. Phys.* **23**, 9–18 (2001). doi: 10.1016/S1350-4533(01)00011-X; pmid: 11344003
54. G. Shin *et al*., Flexible near-field wireless optoelectronics as subdermal implants for broad applications in optogenetics. *Neuron* **93**, 509–521.e3 (2017). doi: 10.1016/j.neuron.2016.12.031; pmid: 28132830
55. H. Lyu *et al*., Synchronized biventricular heart pacing in a closed-chest porcine model based on wirelessly powered leadless pacemakers. *Sci. Rep.* **10**, 2067 (2020). doi: 10.1038/s41598-020-59017-z; pmid: 32034237
56. G. L. Barbruni, P. M. Ros, D. Demarchi, S. Carrara, D. Ghezzi, Miniaturised wireless power transfer systems for neurostimulation: A review. *IEEE Trans. Biomed. Circuits Syst.* **14**, 1160–1178 (2020). doi: 10.1109/TBCAS.2020.3038599; pmid: 33201828
57. D. K. Freeman *et al*., A sub-millimeter, inductively powered neural stimulator. *Front. Neurosci.* **11**, 659 (2017). doi: 10.3389/fnins.2017.00659; pmid: 29230164
58. M. Zargham, P. G. Gulak, Fully integrated on-chip coil in 0.13 μm CMOS for wireless power transfer through biological media. *IEEE Trans. Biomed. Circuits Syst.* **9**, 259–271 (2015). doi: 10.1109/TBCAS.2014.2328318; pmid: 25099630
59. K. L. Montgomery *et al*., Wirelessly powered, fully internal optogenetics for brain, spinal and peripheral circuits in mice. *Nat. Methods* **12**, 969–974 (2015). doi: 10.1038/nmeth.3536; pmid: 26280330
60. T. Nan *et al*., Acoustically actuated ultra-compact NEMS magnetoelectric antennas. *Nat. Commun.* **8**, 296 (2017). doi: 10.1038/s41467-017-00343-8; pmid: 28831042
61. S. Sonmezoglu, J. R. Fineman, E. Maltepe, M. M. Maharbiz, Monitoring deep-tissue oxygenation with a millimeter-scale ultrasonic implant. *Nat. Biotechnol.* **39**, 855–864 (2021). doi: 10.1038/s41587-021-00866-y; pmid: 33782610
62. D. K. Piech *et al*., A wireless millimetre-scale implantable neural stimulator with ultrasonically powered bidirectional communication. *Nat. Biomed. Eng.* **4**, 207–222 (2020). doi: 10.1038/s41551-020-0518-9; pmid: 32076132
63. K. H. Lee *et al*., Ultrasound-driven two-dimensional $Ti_3C_2T_x$ MXene hydrogel generator. *ACS Nano* **14**, 3199–3207 (2020). doi: 10.1021/acsnano.9b08462; pmid: 32078295
64. A. Khalifa *et al*., The Microbead: A 0.009 mm 3 implantable wireless neural stimulator. *IEEE Trans. Biomed. Circuits Syst.* **13**, 971–985 (2019). doi: 10.1109/TBCAS.2019.2939014; pmid: 31484132
65. W. S. Kim *et al*., Organ-specific, multimodal, wireless optoelectronics for high-throughput phenotyping of peripheral neural pathways. *Nat. Commun.* **12**, 157 (2021). doi: 10.1038/s41467-020-20421-8; pmid: 33420038

66. A. Singer *et al*., Magnetoelectric materials for miniature, wireless neural stimulation at therapeutic frequencies. *Neuron* **107**, 631–643.e5 (2020). doi: 10.1016/j.neuron.2020.05.019; pmid: 32516574
67. J. C. Chen *et al*., A wireless millimetric magnetoelectric implant for the endovascular stimulation of peripheral nerves. *Nat. Biomed. Eng.* **6**, 706–716 (2022). doi: 10.1038/s41551-022-00873-7; pmid: 35361934
68. Z. Yu *et al*., Magnetoelectric biol.-implants powered and programmed by a single transmitter for coordinated multisite stimulation. *IEEE J. Solid-State Circuits* **57**, 818–830 (2022). doi: 10.1109/JSSC.2021.3129993; pmid: 36275505
69. Z. Yu *et al*., MagNI: A magnetoelectrically powered and controlled wireless neurostimulating implant. *IEEE Trans. Biomed. Circuits Syst.* **14**, 1241–1252 (2020). doi: 10.1109/TBCAS.2020.3037862; pmid: 33180732
70. W. Kim *et al*., Magnetoelectrics enables large power delivery to mm-sized wireless bioelectronics. *J. Appl. Phys.* **134**, 094103 (2023). doi: 10.1063/5.0156015; pmid: 37692260
71. M. Silverå Ejneby *et al*., Chronic electrical stimulation of peripheral nerves via deep-red light transduced by an implanted organic photocapacitor. *Nat. Biomed. Eng.* **6**, 741–753 (2022). doi: 10.1038/s41551-021-00817-7; pmid: 34916610
72. J. Charthad *et al*., A mm-sized wireless implantable device for electrical stimulation of peripheral nerves. *IEEE Trans. Biomed. Circuits Syst.* **12**, 257–270 (2018). doi: 10.1109/TBCAS.2018.2799623; pmid: 29578414
73. L. Lu *et al*., Biodegradable monocrystalline silicon photovoltaic microcells as power supplies for transient biomedical implants. *Adv. Energy Mater.* **8**, 1703035 (2018). doi: 10.1002/aenm.201703035
74. A. Saha, S. Iqbal, M. Karmaker, S. F. Zinnat, M. T. Ali, A wireless optical power system for medical implants using low power near-IR laser. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2017**, 1978–1981 (2017). doi: 10.1109/EMBC.2017.8037238; pmid: 29060282
75. A. Ahnood *et al*., Diamond encapsulated photovoltaics for transdermal power delivery. *Biosens. Bioelectron.* **77**, 589–597 (2016). doi: 10.1016/j.bios.2015.10.022; pmid: 26476599
76. K. Song, J. H. Han, H. C. Yang, K. I. Nam, J. Lee, Generation of electrical power under human skin by subdermal solar cell arrays for implantable bioelectronic devices. *Biosens. Bioelectron.* **92**, 364–371 (2017). doi: 10.1016/j.bios.2016.10.095; pmid: 27836601
77. D. R. Agrawal *et al*., Conformal phased surfaces for wireless powering of bioelectronic microdevices. *Nat. Biomed. Eng.* **1**, 0043 (2017). doi: 10.1038/s41551-017-0043; pmid: 29226018
78. J. S. Ho *et al*., Wireless power transfer to deep-tissue microimplants. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7974–7979 (2014). doi: 10.1073/pnas.1403002111; pmid: 24843161
79. A. Aldaoud *et al*., Near-field wireless power transfer to stent-based biomedical implants. *IEEE J. Electromagn. RF Microw. Med. Biol.* **2**, 193–200 (2018). doi: 10.1109/JERM.2018.2833386
80. R. Jegadeesan, K. Agarwal, Y.-X. Guo, S.-C. Yen, N. V. Thakor, Wireless power delivery to flexible subcutaneous implants using capacitive coupling. *IEEE Trans. Microw. Theory Tech.* **65**, 280–292 (2017). doi: 10.1109/TMTT.2016.2615623
81. R. Erfani, F. Marefat, A. M. Sodagar, P. Mohseni, Modeling and experimental validation of a capacitive link for wireless power transfer to biomedical implants. *IEEE Trans. Circuits Syst. II Express Briefs* **65**, 923–927 (2018). doi: 10.1109/TCSII.2017.2737140
82. Y. Zhang *et al*., Battery-free, lightweight, injectable microsystem for *in vivo* wireless pharmacology and optogenetics. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21427–21437 (2019). doi: 10.1073/pnas.1909850116; pmid: 31601737
83. Y. Zhang *et al*., Battery-free, fully implantable optofluidic cuff system for wireless optogenetic and pharmacological neuromodulation of peripheral nerves. *Sci. Adv.* **5**, eaaw5296 (2019). doi: 10.1126/sciadv.aaw5296; pmid: 31281895
84. P. Gutruf *et al*., Fully implantable optoelectronic systems for battery-free, multimodal operation in neuroscience research. *Nat. Electron.* **1**, 652–660 (2018). doi: 10.1038/s41928-018-0175-0
85. J. Lim *et al*., A light-tolerant wireless neural recording IC for motor prediction with near-infrared-based power and data telemetry. *IEEE J. Solid-State Circuits* **57**, 1061–1074 (2022). doi: 10.1109/JSSC.2022.3141688; pmid: 36186085
86. J. Kim *et al*., A soft and transparent contact lens for the wireless quantitative monitoring of intraocular pressure. *Nat. Biomed. Eng.* **5**, 772–782 (2021). doi: 10.1038/s41551-021-00719-8; pmid: 33941897
87. J. Lee *et al*., Neural recording and stimulation using wireless networks of microimplants. *Nat. Electron.* **4**, 604–614 (2021). doi: 10.1038/s41928-021-00631-8
88. K. Sahasrabuddhe *et al*., The Argo: A high channel count recording system for neural recording *in vivo*. *J. Neural Eng.* **18**, 015002 (2020). doi: 10.1088/1741-2552/abd0ce; pmid: 33624614
89. J. T. Costello *et al*., A low-power communication scheme for wireless, 1000 channel brain-machine interfaces. *J. Neural Eng.* **19**, 036037 (2022). doi: 10.1088/1741-2552/ac7352; pmid: 35613546
90. R. R. Harrison *et al*., A Low-Power Integrated Circuit for a Wireless 100-Electrode Neural Recording System. *IEEE J. Solid-State Circuits* **42**, 123–133 (2007). doi: 10.1109/JSSC.2006.886567
91. S. R. Nason *et al*., A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain-machine interfaces. *Nat. Biomed. Eng.* **4**, 973–983 (2020). doi: 10.1038/s41551-020-0591-0; pmid: 32719512
92. R. Muller *et al*., A Minimally Invasive 64-Channel Wireless μECoG Implant. *IEEE J. Solid-State Circuits* **50**, 344–359 (2015). doi: 10.1109/JSSC.2014.2364824
93. C. Shi *et al*., Application of a sub-0.1-mm^3 implantable mote for in vivo real-time wireless temperature sensing. *Sci. Adv.* **7**, eabf6312 (2021). doi: 10.1126/sciadv.abf6312; pmid: 33962948
94. Z. Yu *et al*., "Magnetoelectric backscatter communication for millimeter-sized wireless biomedical implants," in *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, Sydney, NSW, Australia, 17–21 October 2022* (ACM, 2022); doi: 10.1145/3495243.3560541
95. M. M. Ghanbari *et al*., A sub-mm3 ultrasonic free-floating implant for multi-mote neural recording. *IEEE J. Solid-State Circuits* **54**, 3017–3030 (2019). doi: 10.1109/JSSC.2019.2936303
96. D. A. Borton, M. Yin, J. Aceros, A. Nurmikko, An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. *J. Neural Eng.* **10**, 026010 (2013). doi: 10.1088/1741-2560/10/2/026010; pmid: 23428937
97. M. Monge, A. Lee-Gosselin, M. G. Shapiro, A. Emami, Localization of microscale devices *in vivo* using addressable transmitters operated as magnetic spins. *Nat. Biomed. Eng.* **1**, 736–744 (2017). doi: 10.1038/s41551-017-0129-2; pmid: 31015668
98. S. Baltsavias *et al*., *In vivo* wireless sensors for gut microbiome redox monitoring. *IEEE Trans. Biomed. Eng.* **67**, 1821–1830 (2020). pmid: 31634824
99. M. S. Wegmueller, M. Oberle, N. Felber, N. Kuster, W. Fichtner, Signal transmission by galvanic coupling through the human body. *IEEE Trans. Instrum. Meas.* **59**, 963–969 (2010). doi: 10.1109/TIM.2009.2031449
100. S. Sriram, S. Avlani, M. P. Ward, S. Sen, Electro-quasistatic animal body communication for untethered rodent biopotential recording. *Sci. Rep.* **11**, 3307 (2021). doi: 10.1038/s41598-021-81108-8; pmid: 33558552
101. B. Chatterjee *et al*., "A 1.15 μW 5.54 mm 3 implant with a bidirectional neural sensor and stimulator SoC utilizing bi-phasic quasi-static brain communication achieving 6kbps-10Mbps uplink with compressive sensing and RO-PUF based collision avoidance," in *2021 Symposium on VLSI Circuits, Kyoto, Japan, 13–19 June 2021* (IEEE, 2021); doi: 10.23919/VLSICircuits52068.2021.9492445
102. Z. Zhao, G. D. Spyropoulos, C. Cea, J. N. Gelinas, D. Khodagholy, Ionic communication for implantable bioelectronics. *Sci. Adv.* **8**, eabm7851 (2022). doi: 10.1126/sciadv.abm7851; pmid: 35385298
103. J. P. Wright *et al*., A fully implantable wireless bidirectional neuromodulation system for mice. *Biosens. Bioelectron.* **200**, 113886 (2022). doi: 10.1016/j.bios.2021.113886; pmid: 34995836
104. A. D. Mickle *et al*., A wireless closed-loop system for optogenetic peripheral neuromodulation. *Nature* **565**, 361–365 (2019). doi: 10.1038/s41586-018-0823-6; pmid: 30602791
105. C. S. Mestais *et al*., WIMAGINE: Wireless 64-channel ECoG recording implant for long term clinical applications. *IEEE Trans. Neural Syst. Rehabil. Eng.* **23**, 10–21 (2015). doi: 10.1109/TNSRE.2014.2333541; pmid: 25014960
106. Y. Jia *et al*., A trimodal wireless implantable neural interface system-on-chip. *IEEE Trans. Biomed. Circuits Syst.* **14**, 1207–1217 (2020). doi: 10.1109/TBCAS.2020.3037452; pmid: 33180731
107. B. Lee *et al*., An implantable peripheral nerve recording and stimulation system for experiments on freely moving animal subjects. *Sci. Rep.* **8**, 6115 (2018). doi: 10.1038/s41598-018-24465-1; pmid: 29666407
108. J. Ausra *et al*., Wireless battery free fully implantable multimodal recording and neuromodulation tools for songbirds. *Nat. Commun.* **12**, 1968 (2021). doi: 10.1038/s41467-021-22138-8; pmid: 33785751
109. Data for: V. Nair, A. N. Dalrymple, Z. Yu, G. Balakrishnan, C. J. Bettinger, D. J. Weber, K. Yang, J. T. Robinson, Miniature battery-free bioelectronics, Zenodo (2023); doi: 10.5281/zenodo.8310306
110. V. Nair, B. Tian, "Perspectives for seamless integration of bioelectronic systems in neuromedicine," in *Neural Interface Engineering*, L. Guo, Ed. (Springer, 2020); pp. 365–381; http://link.springer.com/10.1007/978-3-030-41854-0_16.
111. C. Lee *et al*., "A miniaturized wireless neural implant with body-coupled data transmission and power delivery for freely behaving animals," in *2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 20–26 February 2022* (IEEE, 2022); doi: 10.1109/ISSCC42614.2022.9731733
112. O. Padron *et al*., PD66-01 treatment of urinary urgency incontinence (UUI) with an ultra-miniaturized sacral nerve modulation (snm) system: Preliminary outcomes of the SANS-UUI study. *J. Urol.* **206**, e1158 (2021). doi: 10.1097/JU.0000000000002110.01
113. D. Seo *et al*., Wireless recording in the peripheral nervous system with ultrasonic neural dust. *Neuron* **91**, 529–539 (2016). doi: 10.1016/j.neuron.2016.06.034; pmid: 27497221
114. K. Patch, Neural dust swept up in latest leap for bioelectronic medicine. *Nat. Biotechnol.* **39**, 255–256 (2021). doi: 10.1038/s41587-021-00856-0; pmid: 33692519
115. T. Okabe *et al*., Leadless left ventricular stimulation with WiSE-CRT System - Initial experience and results from phase I of SOLVE-CRT Study (nonrandomized, roll-in phase). *Heart Rhythm* **19**, 22–29 (2022). doi: 10.1016/j.hrthm.2021.06.1195; pmid: 34332966
116. P. Ritter *et al*., Early performance of a miniaturized leadless cardiac pacemaker: The Micra Transcatheter Pacing Study. *Eur. Heart J.* **36**, 2510–2519 (2015). doi: 10.1093/eurheartj/ehv214; pmid: 26045305
117. G. Yao *et al*., Effective weight control via an implanted self-powered vagus nerve stimulation device. *Nat. Commun.* **9**, 5349 (2018). doi: 10.1038/s41467-018-07764-z; pmid: 30559435
118. H. Ouyang *et al*., Symbiotic cardiac pacemaker. *Nat. Commun.* **10**, 1821 (2019). doi: 10.1038/s41467-019-09851-1; pmid: 31015519
119. H. Ryu *et al*., Self-rechargeable cardiac pacemaker system with triboelectric nanogenerators. *Nat. Commun.* **12**, 4374 (2021). doi: 10.1038/s41467-021-24417-w; pmid: 34272375
120. C. Dagdeviren *et al*., Conformal piezoelectric energy harvesting and storage from motions of the heart, lung, and diaphragm. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 1927–1932 (2014). doi: 10.1073/pnas.1317233111; pmid: 24449853
121. W. Ren *et al*., High-performance wearable thermoelectric generator with self-healing, recycling, and Lego-like reconfiguring capabilities. *Sci. Adv.* **7**, eabe0586 (2021). doi: 10.1126/sciadv.abe0586; pmid: 33568483
122. S. Hong *et al*., Wearable thermoelectrics for personalized thermoregulation. *Sci. Adv.* **5**, eaaw0536 (2019). doi: 10.1126/sciadv.aaw0536; pmid: 31114803
123. K. Nan *et al*., Compliant and stretchable thermoelectric coils for energy harvesting in miniature flexible devices. *Sci. Adv.* **4**, eaau5849 (2018). doi: 10.1126/sciadv.aau5849; pmid: 30406207
124. C. S. Kim *et al*., Self-powered wearable electrocardiography using a wearable thermoelectric power generator. *ACS Energy Lett.* **3**, 501–507 (2018). doi: 10.1021/acsenergylett.7b01237

ACKNOWLEDGMENTS

V.N. thanks A. Singer, J. Woods, W. Kim, A. Tuppen, and G. Duret for useful scientific discussions. **Funding:** This work received no funding. **Author contributions:** V.N. and J.T.R. designed the concept, wrote the introduction and the sections on battery-free

concepts and WPT, and prepared Figs. 1 to 3. G.B. and C.J.B. wrote the section on energy harvesting. Z.Y. and K.Y. wrote the section on communications and prepared Fig. 4. A.N.D. and D.J.W. wrote the sections on wireless networks and current technologies. All authors reviewed the manuscript and provided input. **Competing interests:** J.T.R. is a co-founder and CEO of Motif Neurotech, Inc., and holds an equity stake in and receives financial compensation from Motif Neurotech, Inc. K.Y. holds equity in Motif Neurotech, Inc. J.T.R. and K.Y. are inventors on patent application number PCT/US2020/026688 submitted by Rice University that covers magnetoelectric power and data transfer to bioelectronic implants. The remaining authors declare no competing interests. **Data and materials availability:** Datasets used in this study are available on Zenodo (*109*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

Submitted 10 June 2022; accepted 28 September 2023
10.1126/science.abn4732

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors

Melissa L. Boby†, Daren Fearon†, Matteo Ferla†, Mihajlo Filep†, Lizbé Koekemoer†, Matthew C. Robinson†, The COVID Moonshot Consortium, John D. Chodera*, Alpha A. Lee*, Nir London*, Annette von Delft*, Frank von Delft*

INTRODUCTION: COVID-19 became a global pandemic partially as a result of the lack of easily deployable, broad-spectrum oral antivirals, which complicated its containment. Even endemically, and with effective vaccinations, it will continue to cause acute disease, death, and long-term sequelae globally unless there are accessible treatments. COVID-19 is not an isolated event but instead is the latest example of a viral pandemic threat to human health. Therefore, antiviral discovery and development should be a key pillar of pandemic preparedness efforts.

RATIONALE: One route to accelerate antiviral drug discovery is the establishment of open knowledge bases, the development of effective technology infrastructures, and the discovery of multiple potent antivirals suitable as starting points for the development of therapeutics. In this work, we report the results of the COVID Moonshot—a fully open science, crowdsourced, and structure-enabled drug discovery campaign—against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) main protease (Mpro). This collaboration may serve as a roadmap for the potential development of future antivirals.

RESULTS: On the basis of the results of a crystallographic fragment screen, we crowdsourced design ideas to progress from fragment to lead compounds. The crowdsourcing strategy yielded several key compounds along the optimization trajectory, including the starting compound of what became the primary lead series. Three additional chemically distinct lead series were also explored, spanning a diversity of chemotypes.

The collaborative and highly automated nature of the COVID Moonshot Consortium resulted in >18,000 compound designs, >2400 synthesized compounds, >490 ligand-bound x-ray structures, >22,000 alchemical free-energy calculations, and >10,000 biochemical measurements—all of which were made publicly available in real time. The recently approved antiviral ensitrelvir was identified in part based on crystallographic data from the COVID Moonshot Consortium.

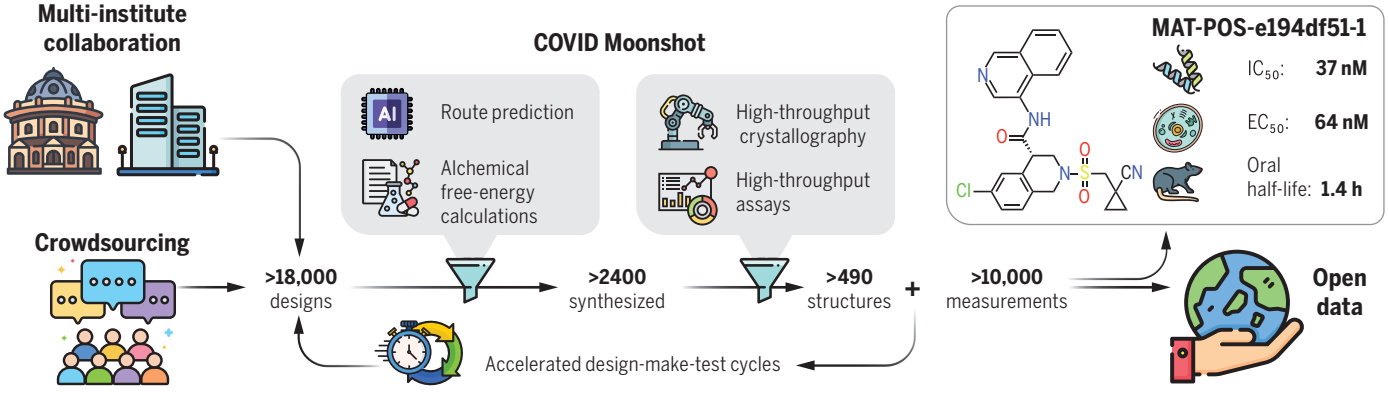
This campaign led to the discovery of a potent [median inhibitory concentration (IC_{50}) = 37 ± 2 nM] and differentiated (noncovalent and nonpeptidic) lead compound that also exhibited potent cellular activity, with a median effective concentration (EC_{50}) of 64 nM in A549-ACE2-TMPRSS2 cells and 126 nM in HeLa-ACE2 cells without measurable cytotoxicity. Although the pharmacokinetics of the reported compound is not yet optimal for therapeutic development, it is a promising starting point for further antiviral discovery and development.

CONCLUSION: The success of the COVID M… shot project in producing potent antivir…, building open knowledge bases, accelerating external discovery efforts, and functioning as a useful information-exchange hub is an example of the potential effectiveness of open science antiviral discovery programs. The open science, patent-free nature of the project enabled a large number of collaborators to provide in-kind support, including synthesis, assays, and in vitro and in vivo experiments. By making all data immediately available and ensuring that all compounds are purchasable from Enamine without the need for materials transfer agreements, we aim to accelerate research globally along parallel tracks. In the process, we generated a detailed map of the structural plasticity of Mpro, extensive structure-activity relationships for multiple chemotypes, and a wealth of biochemical activity data to spur further research into antivirals and discovery methodologies. We hope that this can serve as an alternative model for antiviral discovery and future pandemic preparedness.

Further, the project also showcases the role of machine learning, computational chemistry, and high-throughput structural biology as force multipliers in drug design. Artificial intelligence and machine learning algorithms help accelerate chemical synthesis while balancing multiple competing molecular properties. The design-make-test-analyze cycle was accelerated by these algorithms combined with planetary-scale biomolecular simulations of protein-ligand interactions and rapid structure determination. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: john.chodera@choderalab.org (J.D.C.); alpha.lee@postera.ai (A.A.L.); nir.london@weizmann.ac.il (N.L.); annette.vondelft@cmd.ox.ac.uk (A.v.D.); frank.von-delft@diamond.ac.uk (F.v.D.)
†These authors contributed equally to this work.
Cite this article as M. L. Boby *et al.*, *Science* **382**, eabo7201 (2023). DOI: 10.1126/science.abo7201

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abo7201



The COVID Moonshot Consortium. An open science, crowdsourced drug discovery campaign against the SARS-CoV-2 Mpro led to a potent, noncovalent, and nonpeptidic inhibitor scaffold with lead-like properties. We generated copious structural, biochemical, and pharmacological data that were shared rapidly and openly, creating a rich, open, and intellectual property–free knowledge base for future anticoronavirus drug discovery.

CREDIT: ICONS MADE BY FREEPIK AND GOOD WARE FROM WWW.FLATICON.COM

RESEARCH ARTICLE

CORONAVIRUS

Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors

Melissa L. Boby[1,2,3]†, Daren Fearon[4,5]†, Matteo Ferla[6]†, Mihajlo Filep[7]†, Lizbé Koekemoer[8,9]†, Matthew C. Robinson[10]†, The COVID Moonshot Consortium‡, John D. Chodera[3]*, Alpha A. Lee[10]*, Nir London[7]*, Annette von Delft[6,8,9]*, Frank von Delft[4,5,8,9,11]*

We report the results of the COVID Moonshot, a fully open-science, crowdsourced, and structure-enabled drug discovery campaign targeting the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) main protease. We discovered a noncovalent, nonpeptidic inhibitor scaffold with lead-like properties that is differentiated from current main protease inhibitors. Our approach leveraged crowdsourcing, machine learning, exascale molecular simulations, and high-throughput structural biology and chemistry. We generated a detailed map of the structural plasticity of the SARS-CoV-2 main protease, extensive structure-activity relationships for multiple chemotypes, and a wealth of biochemical activity data. All compound designs (>18,000 designs), crystallographic data (>490 ligand-bound x-ray structures), assay data (>10,000 measurements), and synthesized molecules (>2400 compounds) for this campaign were shared rapidly and openly, creating a rich, open, and intellectual property–free knowledge base for future anticoronavirus drug discovery.

Despite rapid progress in vaccine development, the global failure to abate COVID-19, which culminated in more than 690 million confirmed cases worldwide by July 2023, will likely cause the virus to become endemic (*1*) and continue to cause a significant number of deaths, especially in the Global South, unless there is an accessible treatment (*2*). Antiviral therapeutics are a necessary and complementary strategy to vaccination to control COVID-19 (*3*). Several directly acting oral antivirals are now approved for use against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, including ritonavir-boosted nirmatrelvir (*4*), ensitrelvir (Japan) (*5*), and molnupiravir (*6*).

[1]Pharmacology Graduate Program, Weill Cornell Graduate School of Medical Sciences, New York, NY 10065, USA. [2]Program in Chemical Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [3]Program in Computational and Systems Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [4]Diamond Light Source Ltd., Harwell Science and Innovation Campus, Didcot, UK. [5]Research Complex at Harwell, Harwell Science and Innovation Campus, Didcot, UK. [6]Oxford Biomedical Research Centre, National Institute for Health Research, University of Oxford, Oxford, UK. [7]Department of Chemical and Structural Biology, The Weizmann Institute of Science, Rehovot, Israel. [8]Centre for Medicines Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [9]Structural Genomics Consortium, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [10]PostEra Inc., Cambridge, MA 02142, USA. [11]Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa.
***Corresponding author. Email: john.chodera@choderalab.org (J.D.C.); alpha.lee@postera.ai (A.A.L.); nir.london@weizmann.ac.il (N.L.); annette.vondelft@cmd.ox.ac.uk (A.v.D.); frank.von-delft@diamond.ac.uk (F.v.D.)**
†These authors contributed equally to this work.
‡The COVID Moonshot Consortium authors and affiliations are listed at the end of this paper.

COVID-19 is not an isolated event but rather the latest exemplar of a series of threats to human health caused by beta-coronaviruses, which have also been responsible for the SARS (2003) and Middle East respiratory syndrome (MERS) (2010) pandemics (*7*). Open knowledge bases and technology infrastructures for antiviral drug discovery will enable pandemic preparedness by refreshing the antivirals pipeline and providing multiple starting points for the development of therapeutics. In this work, we report the open science discovery of a potent SARS-CoV-2 antiviral lead compound and a roadmap for the potential development of future SARS-CoV-2 and pan-coronavirus antivirals.

The SARS-CoV-2 main protease (Mpro; or 3CL-protease) is an attractive target for antiviral development because of its essential role in viral replication, a large degree of conservation across coronaviruses, and dissimilarity of its structure and substrate profile to human proteases (*8*) (fig. S1). Pioneering studies during and after the 2003 SARS pandemic established the linkage between Mpro inhibition and antiviral activity in cell culture (*9*). This work has been corroborated by in vitro and in vivo studies for SARS-CoV-2 (*10*, *11*) and the clinical success of nirmatrelvir (the Mpro inhibitor component of Paxlovid) (*12*) and ensitrevir (Xocova) (*13*, *14*).

To warrant early use in the course of disease or even prophylactically among at-risk populations, an antiviral drug would need to be orally available with an excellent safety profile. Given the historical difficulties in developing peptidomimetic compounds into oral drugs and the risk of downstream idiosyncratic hazards of covalent inhibition, we chose to pursue noncovalent, nonpeptidomimetic scaffolds. First-generation oral Mpro inhibitors have now demonstrated clinical efficacy (*15*, *16*), but the need for cytochrome P450 3A4 (CYP3A4) inhibitor co-dosing (ritonavir, in the case of Paxlovid) to achieve sufficient human exposure may substantially limit use in at-risk populations because of drug-drug interactions (*17*). There remains a need for chemically differentiated oral antiviral protease inhibitors with the potential to enter clinical development.

Crowdsourced progression of x-ray fragment hits rapidly generated potent lead compounds with diverse chemotypes

The COVID Moonshot is an open science drug discovery campaign targeting SARS-CoV-2 Mpro (*18*, *19*), building off a rapid crystallographic fragment screening campaign that assessed >1250 unique fragment-soaked crystals screened within weeks to identify 71 hits that densely populated the active site (Fig. 1A) (*20*). This dataset was posted online on 18 Mar 2020 (*21*), days after the screen was completed (*21*). The noncovalent fragment hits did not show detectable inhibition in a fluorescence-based enzyme activity assay [assay dynamic range median inhibitory concentration (IC_{50}) < 100 μM]. However, they provided a high-resolution map of key interactions that optimized compounds may exploit to inhibit Mpro (*22*).

Numerous approaches have been proposed to advance from fragments to lead compounds (*23*, *24*). One strategy, fragment merging, aims to combine multiple fragments into a single, more-potent molecule, whereas fragment expansion elaborates a fragment to engage neighboring interactions. Although these strategies are usually applied to a single fragment or a handful of fragments, our large-scale fragment screen produced a dense ensemble of hits, which provided an opportunity for rapid lead generation by combining chemotypes from multiple fragments. Nonetheless, this approach requires heuristic chemical reasoning that accounts for the spatial orientation of fragments in the binding site—a feat that can challenge algorithms but is also potentially solvable by humans. Building on successes in crowdsourced protein (*25*) and RNA (*26*) design campaigns, we hypothesized that crowdsourced human analysis and algorithmic strategies could accelerate the generation of potent lead compounds and furnish diverse chemical matter because different chemists would use different approaches and reasoning strategies.

We launched an online crowdsourcing platform (https://postera.ai/covid) on 18 March 2020 (Fig. 1B), soliciting participants to submit compounds designed on the basis of the fragment hits (*19*). Compounds selected for synthesis were evaluated by biochemical assays (data S1) and x-ray crystallography, and the results were released rapidly on the same platform, which enabled contributing designers to build on all available data as well as on designs contributed

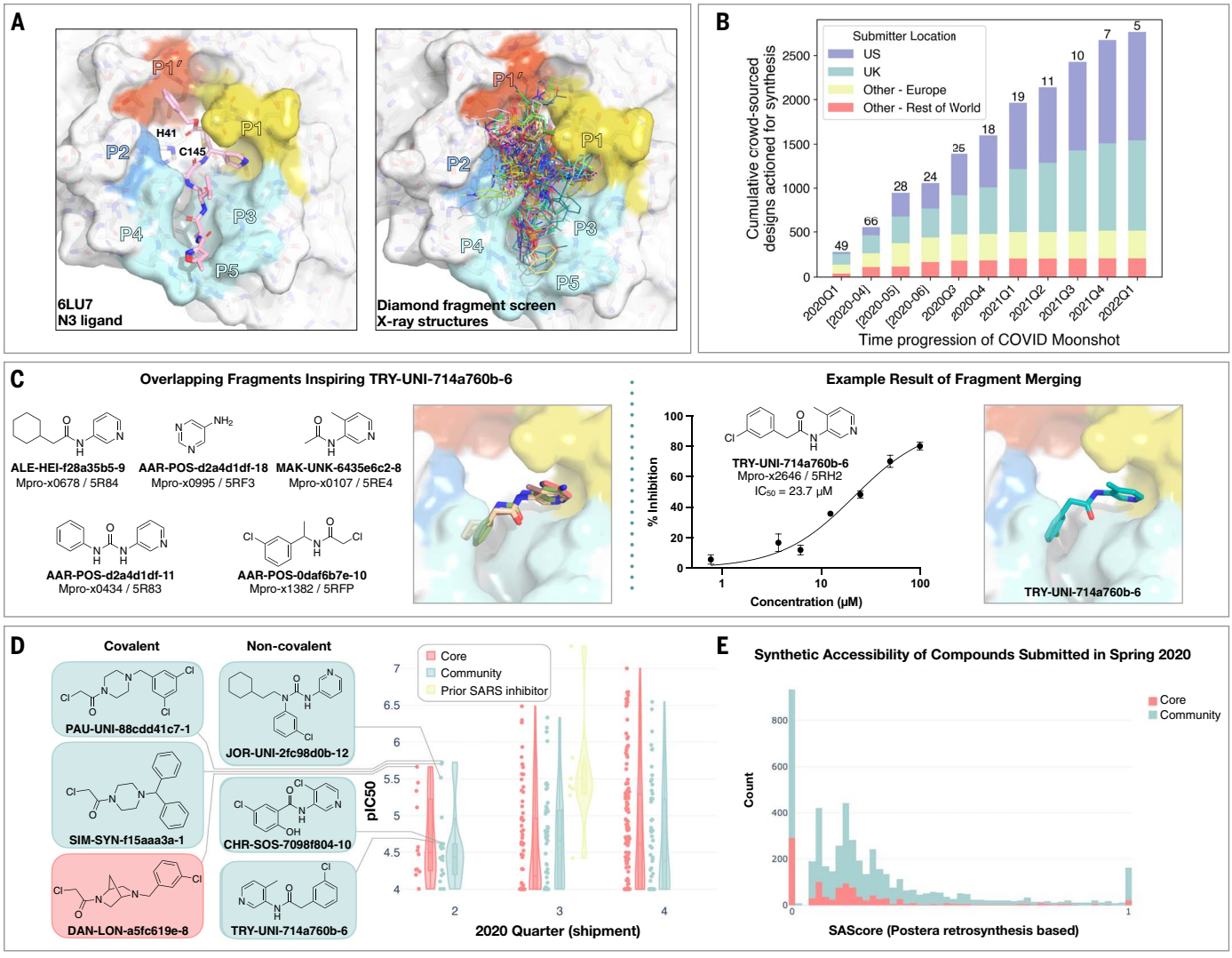


Fig. 1. Crowdsourcing rapidly identified chemotype scaffolds by merging fragment hits. (**A**) A Diamond/XChem fragment screen that initiated this SARS-CoV-2 Mpro inhibitor discovery campaign generated 71 hits that completely cover the Mpro active site, with a variety of chemotypes engaging each pocket; 1638 x-ray datasets were collected, and 96 solved structures for hits were publicly posted (*20*). The peptidomimetic N3 ligand is shown on the left for comparison to indicate natural substrate engagement in the binding site, defining the peptide sidechain numbering scheme used throughout this work. The nucleophilic Cys^{145} reacts with the scissile peptide bond between P1 and P1'; His^{41}-Cys^{145} form a catalytic dyad whose coupled charge states shuttle between zwitterionic and neutral states (*90*). (**B**) On 18 March 2020, the COVID Moonshot set up a crowdsourcing website to empower scientists across the globe to contribute molecule designs. The number of designs actioned for synthesis each quarter (except for the 2020 Q2, which is shown per-month in brackets) is shown, subdivided by the region of the submitter of the design idea. The total number of unique submitters that contributed actioned designs for that quarter is shown on top of the bars. (**C**) Many submissions, such as TRY-UNI-714a760b-6, exploited spatially overlapping fragment hits to design potent leads that are synthetically facile. (**D**) Experimental biochemical potency of designs broken down by submission group. Multiple submissions in 2020 from the community were more potent than the best designs from the core team, as seen for the top three chloroacetamide structures (left) and noncovalent structures (right). (**E**) Distribution of synthetic accessibility scores (SAScores) for designs contributed by the core team and the community. The concern that community submissions may be of poor quality is not supported by the fact that these were as synthetically accessible as those designed by the core team (median: community, 0.17; core, 0.13). Half of the outliers (SAScore = 1) were primarily natural products, which are hard to achieve through organic chemistry.

by others. To facilitate transparency and maximal speed, and to avoid delays around intellectual property (IP), all designers were asked to contribute their designs directly into the public domain, with every design and all related experimental data immediately disclosed online and made openly available, explicitly free of IP restrictions. This aggressive open science policy enabled contributors from multiple fields in both academia and industry to freely share their ideas. Within the first week, we received more than 2000 submissions, representing a diverse set of design strategies (data S2).

Many submissions exploited spatially overlapping fragment hits. For example, the submission TRY-UNI-714a760b-6 was inspired by five overlapping fragments, furnishing a noncovalent inhibitor with a SARS-CoV-2 Mpro enzymatic IC_{50} of 23.7 μM (Fig. 1C). This compound seeded the aminopyridine series, whose optimization is described in detail below. Only 11 of the 768 fragments in the DSi-poised library (*27*, *28*) contained a 3-amino pyridine; yet, four of them were successfully identified in the crystallographic fragment screen and were consequently picked up for merging by the designers. Apart from the aminopyridine series,

our campaign identified three other major chemically distinct lead series with measurable potencies against SARS-CoV-2 Mpro inspired by reported SARS-CoV-1 inhibitors (fig. S2). Those compounds span the same binding pocket but feature different chemotypes, and the large quantity of structure-activity relationship (SAR) subsequently generated for these series furnishes multiple backup series with different risk profiles. Other groups have subsequently further elaborated on the Ugi (*29*, *30*) and the benzotriazole series that we generated (*31*).

Analysis of the submissions provides some hints to the utility of crowdsourcing as a general strategy for hit-discovery or hit-to-lead campaigns. A qualitative assessment of the textual description of submitted designs (fig. S3) hints that many of the designers used tools such as docking to assess fragment "linking," "merging," or "combination." When trying to more-thoroughly categorize submissions, it does not appear that hypothesis-driven designs perform better than docking-driven designs; however, "predicting" historical SARS inhibitors is the best-performing strategy (fig. S4 and Fig. 1D). Throughout the campaign, designs were contributed both by the core group of laboratories and medicinal chemists leading this project and by the community. One could hypothesize that the core group being committed to the project, as well as being thoroughly invested in the campaign details, would contribute more-potent designs. However, there is no obvious difference in the distributions of designs produced by the core group versus the community in the early stages of the campaign (Fig. 1D), nor were the designs contributed by the community less synthetically accessible (Fig. 1E). Later in the campaign (lead optimization stage), the number of submissions from the community decreased, and comparing potency became irrelevant as other attributes of the molecules were being optimized. It is important to mention that several key compounds along the optimization trajectory of our lead were contributed by the community and not core group members: TRY-UNI-714a760b-6, ADA-UCB-6c2cb422-1, and VLA-UCB-1dbca3b4-15 (the racemic mixture of MAT-POS-b3e365b9-1). Although anecdotal, this example demonstrates the potential power of crowdsourcing as a strategy to drive fragment-to-lead campaigns.

Technologies to support rapid optimization cycles

With a growing number of chemically diverse submissions, we relied on a team of experienced medicinal chemists supported by computational methods to aid in triaging design proposals with the goal of increasing potency. To execute a rapid fragment-to-lead campaign, we used models to plan synthetic routes, enumerate synthetically accessible virtual libraries, and estimate potencies to prioritize which compounds to target for synthesis. We did not use an "autonomous" approach—expert judgment is used to make decisions given all the model predictions. Furthermore, in the context of a fast-moving campaign, we prioritized making progress over granular "human-versus-machine" evaluations.

Synthetic route predictions guided decision-making to accelerate design-make-test-analyze cycles

We used an established synthetic contract research organization (CRO), Enamine, to carry out rapid synthesis of progressed compound designs. To take full advantage of the available building block collection, we used a machine learning approach that plans efficient retrosynthetic routes to predict synthetic tractability (*32*, *33*). We automatically computed synthetic routes for all crowdsourced submissions using Enamine's in-stock building block inventories. From the computed routes, synthetic complexity was estimated on the basis of the number of steps and the probability of success of each step. The synthetic accessibility score, as well as the predicted synthetic route, were then used to aid medicinal chemistry decision-making. Our predicted synthetic complexity correlated with the actual time taken to synthesize target compounds, and the algorithm was able to pick out advanced intermediates as starting materials (Fig. 2A).

Alchemical free-energy calculations prioritized potent compounds for synthesis

We estimated potency of proposed designs and virtual synthetic libraries of analogs using alchemical free-energy calculations (*34*–*36*), an accurate physical modeling technique that has hitherto not been deployed in a high-throughput setup because of its prohibitive computational cost. We used Folding@home (*37*)—a worldwide distributed computing network where hundreds of thousands of volunteers around the world contributed computing power to create the world's first exascale computing resource (*38*)—to compute the free energy of binding of all 20,000+ crowdsourced and internal design submissions using the Open Force Field Initiative "Parsley" small-molecule force fields (*39*) and nonequilibrium switching with the open source PERSES alchemical free-energy toolkit (*40*–*42*) based on the graphics processing unit (GPU)–accelerated OpenMM framework (*38*, *43*) (see Materials and methods). Comprehensive sampling was prioritized over efficiency of computation given the abundant compute resources available on Folding@home.

We first performed a small retrospective study using bioactivity data generated from the first week of crowdsourced compound designs, triaged solely using synthetic accessibility. The results of these free-energy calculations showed good correlation with experimentally measured affinities (Fig. 2B). Henceforth, alchemical free-energy calculations were used as an additional (though not the sole) criterion to guide compound selection and iterative design (see Data and materials availability statement). During the campaign, distinct objectives were solicited from submitters to address medicinal chemistry problems, and free-energy calculations were used to assess these submissions on the basis of predicted potency. Fig. 2C shows that predicted $-\log_{10}$ IC_{50} (pIC_{50}) tracks experimental measurements across three chronologically distinct design campaigns: decoration of the benzopyran ring, replacement of the benzopyran system, and replacement of the isoquinoline system. Some design ideas with low predicted pIC_{50} were synthesized because the medicinal chemistry team balanced between gaining insights on structure-activity and structure-property relationship and potency optimization. The champion compounds from each design campaign are highlighted in the right panel of Fig. 2C. Although free-energy calculations identified multiple potency-improving transformations, the strategically useful one was the swap from pyran to a piperidine sulfonamide system, which is on the critical path to the lead compound. On average, 80 GPU-hours per compound were used across the three panels (Materials and methods).

A major strength of alchemical free-energy calculations proved to be their ability to select potent analogs from virtual synthetic libraries from which the medicinal chemistry team had already selected compounds sharing a common intermediate as well as highlighting submitted designs predicted to be highly potent but where major synthetic effort would be required. Our design team prioritized for synthesis small libraries suggested by the aforementioned computational approaches. Chemically related groups of outliers frequently provided chemical insight that informed modeling choices (fig. S5). The approach was not without drawbacks, including the need to focus on a single reference compound and structure to design transformation networks (rather than leveraging the abundant structural data), the requirement that protonation states be fixed for the entire calculation (requiring the entire transformation network to be recomputed to assess a different protonation state), and the relatively large computational cost required to handle large replacements (see Materials and methods). The method is also not uniformly accurate across all chemical transformations, and accurately estimating its accuracy beforehand is challenging. For example, isoquinoline replacements show lower correlation between calculated and predicted free energy (Fig. 2B, panel 3) compared with the benzopyran replacements (Fig. 2B, panel 2).

Nanomole-scale high-throughput chemistry enabled rapid evaluation of SAR

A complementary method for rapid SAR evaluation was the use of nanomole-scale high-throughput chemistry (HTC) (*44*, *45*) coupled with a "direct to biology" (*46*–*48*) biochemical

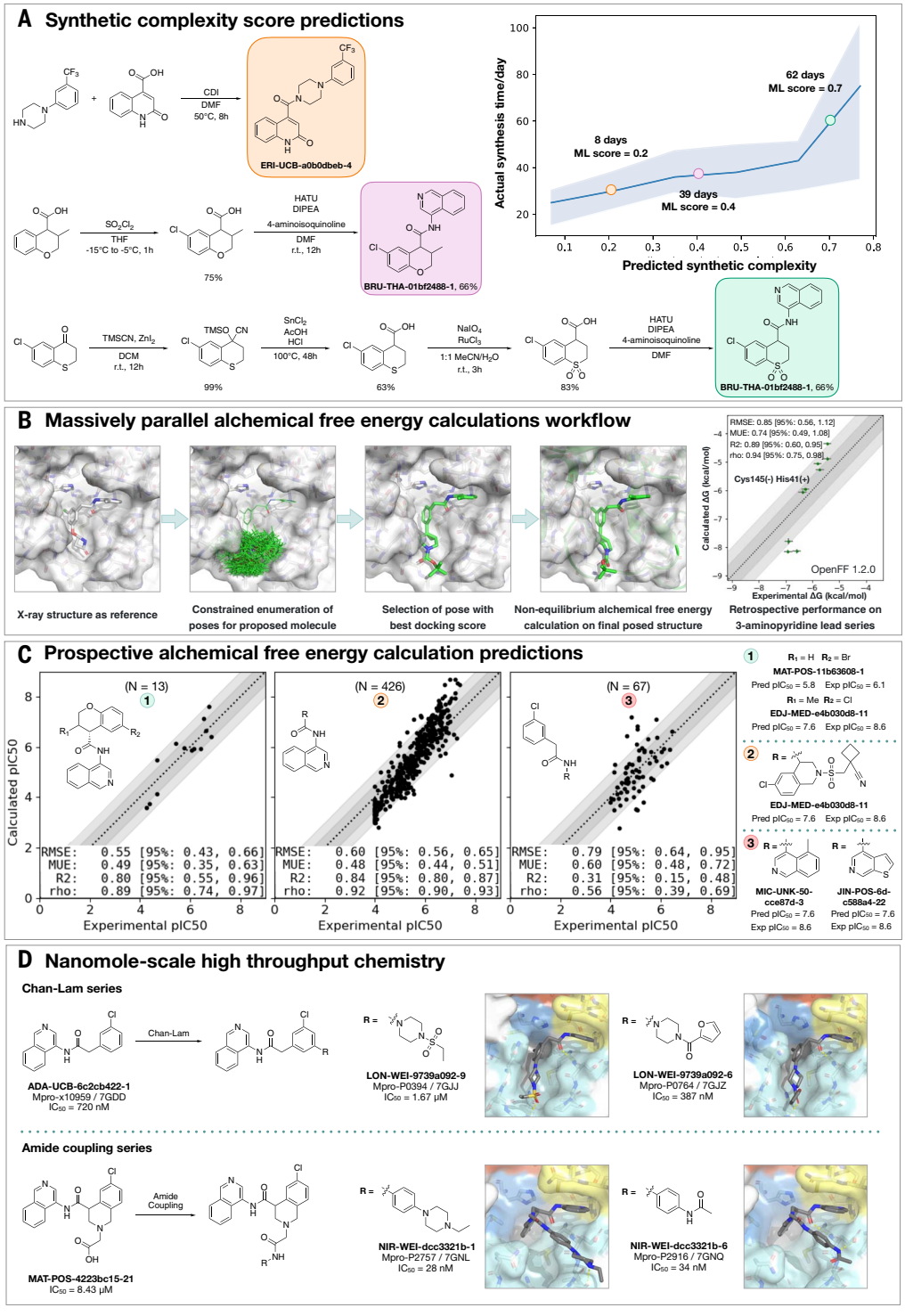


Fig. 2. Strategies to support rapid optimization cycles. (**A**) Machine learning forecasts experimental synthesis time (left) and returns efficient routes that leverage more than 10 million in-stock advanced intermediates (right). Our algorithm predicts the probability of each step being successful and predicts synthetic accessibility by taking the product of the probabilities along the whole route. We analyzed all compounds made in COVID Moonshot from 1 May 2020 to 1 July 2021 (n = 898). The right panel exemplifies the experimental execution of the predicted routes, demonstrating the ability of the algorithm to build on functionalized intermediates to shorten synthesis. (**B**) Applying alchemical free-energy calculations at scale enables us to estimate the potency of compounds. Retrospective assessment of our automated free-energy calculation workflow on early compounds in the 3-aminopyridine series in the first month of the COVID Moonshot campaign suggested that free-energy calculations could provide good predictive utility, which inspired confidence for large-scale deployment during this campaign. Here, the absolute free energy of binding (ΔG) is shown in the rightmost panel by adding a constant offset to the computed relative free-energy differences. (**C**) Alchemical free-energy predictions for all submissions elaborating on the depicted scaffold for three representative batches of prospective free-energy calculations plotted as calculated (converted using Cheng-Prusoff equation) versus experimental pIC_{50}. Simulations were run using Mpro in dimer form, with neutral catalytic dyad and no restraints. Each batch (numbered 1 to 3 from left to right) is annotated with its scaffold, and top-scoring candidates are shown on the right-hand side (numbered 1 to 3 from top to bottom)—for these, the structure names are shown together with their predicted and experimental pIC_{50} ("Pred" and "Exp," respectively). Statistical performance with 95% confidence intervals for each batch is shown as a table in each scatterplot. (**D**) Two examples of nanomole-scale HTC campaigns used to optimize the potency of intermediate binders, centering on the Chan-Lam reaction (fig. S7) and amide couplings (fig. S8). Direct biochemical screening of crude reactions identified candidates that were resynthesized and in both cases were able to improve the potency of the parent compound. Soaking of crude reaction mixtures of the most potent biochemical hits into Mpro crystals provided complex structures with the identified hits (Chan-Lam PDBs: 7GJJ/7GJZ, resolution: 1.75Å/1.65Å; Amide coupling PDBs: 7GNL/7GNQ, resolution: 1.68Å/1.53Å). In both cases, new interactions were discovered, explaining the improved activity. Although for the Chan-Lam reaction campaign, the extended compounds occupied the intended P4, for the amide-coupling vector, all compounds extended into the P3/5 pockets.

screening. Two examples include the optimization of the Chan-Lam reaction (*49*) to extend molecule ADA-UCB-6c2cb422-1 and amide coupling to extend MAT-POS-4223bc15-21 (Fig. 2D). In both cases, we determined the cocrystal structures of the parent compounds (fig. S6) and suggested vectors that could target the P4 pocket of Mpro. Optimization of the reaction conditions was performed for the starting building block with model amines (figs. S7 and S8), and the optimal conditions were applied to HTC with a library of 300 amine building

blocks (data S3). Yield estimation was performed in both cases and showed for the Chan-Lam library that only 29 of the library yielded >30% of the desired product compared with 151 for the amide coupling. Nevertheless, the crude mixtures were subjected to a biochemical assay against Mpro (data S3). Seven compounds were selected for resynthesis from the Chan-Lam series and 20 from the amide series (fig. S9). In parallel to synthesis, the crude reaction mixtures were subjected to soaking and x-ray crystallography. The structures verified that the extended compounds do adopt a similar binding mode to the parent. Chan-Lam–extended compounds occupied P4, whereas the amides extended toward P3/P5, in both cases forming new interactions with Mpro (Fig. 2D). Upon resynthesis, one of the Chan-Lam compounds was able to slightly improve over the parent compound IC_{50}. Several of the amide-coupling series were able to improve by up to 300-fold on the parent acid-compound (up to threefold on the corresponding methylamide), with the best inhibitor exhibiting an IC_{50} of 28 nM against Mpro.

Covalent targeting strategies

Another approach that was attempted to rapidly gain potency was the use of electrophiles to covalently target the catalytic C145. The original fragment screen (*20*) that launched this effort included electrophiles (*50*) and resulted in 48 structures of covalently bound fragments, the majority of which were chloroacetamides. Some of the earliest, and most-potent, fragment merges explored by both the core group and the community were of chloroacetamide (Fig. 1D), and further optimization improved chloroacetamide fragments' IC_{50} values to as low as 300 nM (fig. S10). Chloroacetamides, however, are not considered suitable for therapeutics, and therefore we aimed to move toward acrylamides by derivatizing potent reversible hits (*30*) (fig. S11). Ultimately, we focused on a noncovalent series, but the chlorophenyl moiety that remained throughout the series was adopted from a chloroacetamide hit fragment (AARPOS-0daf6b7e-10; Fig. 1C).

High-throughput structural biology uncovered binding modes and interactions underlying potency

We selected compounds on the basis of synthetic tractability and alchemical free-energy calculations. We profiled every compound through crystal soaking and x-ray diffraction, totaling in 587 structures (see table S1 and fig. S12 for average statistics, data S4 for crystallographic and refinement statistics, and fig. S13 for ligand density for the structures highlighted in this manuscript). Analysis of a subset of this large trove of structural data (n = 367, up to July 2021) reveals the hotspots for ligand engagement and plasticity of each binding pocket. Fig. 3 highlights the statistics of intermolecular interactions between the residues and our ligands. The P1 and P2 pockets are the hotspots of interactions; yet, the interaction patterns are starkly different. The salient interactions sampled by our ligands in the P1 pocket are H163 (H-bond donor), E166 (H-bond acceptor), and N142 (hydrophobic interactions), whereas P2 interactions are dominated by π-stacking interactions with H41 and hydrophobic interactions with M165. The P1′ and P3/4/5 pockets are sparingly sampled by our ligands; the former can be targeted through hydrophobic interactions (T25) and the latter through H bonds (Q192).

This pattern of intermolecular interactions is reflected in the plasticity of the different subpockets. The dominance of directional interactions in P1 renders it more rigid than P2 (Fig. 4). The rigidity is also dependent on the chemical series (fig. S2), with the Ugi and benzotriazole series being able to deform the P2 pocket. Those series comprise more heavy atoms and span a larger region of the binding site; thus, changes in P2 pocket interactions could be better tolerated.

Design of a SARS-CoV-2 Mpro inhibitor lead series with potent antiviral activity

Our medicinal chemistry strategy was driven by the design of potent ligand-efficient and geometrically compact inhibitors that fit tightly in the substrate binding pocket. The former strategy aimed to increase the probability of achieving oral bioavailability, whereas the latter heuristic was motivated by the substrate envelope hypothesis for avoiding viral resistance (*51*). Fig. 5A outlines the critical intermediates on the path toward an optimized lead compound.

Starting from the fragment hit, we explored the P1 pocket, which admits a steep SAR—perhaps unsurprising given its rigidity and preference for directional H-bond interactions (Fig. 3A). An increase in potency was unlocked by replacing pyridine with isoquinoline, which picks up additional hydrophobic interactions with N142. The SAR around the P2 pocket is considerably more tolerant to modifications and broadly favors hydrophobic moieties. A step-change in potency was achieved by rigidifying the scaffold: We introduced a tetrahydropyran ring to transform the P2 substituent into a chromane moiety (compound MAT-POS-b3e365b9-1; the racemic mixture VLA-UCB-1dbca3b4-15, which was initially synthesized, has a IC_{50} of 360 nM; Fig. 5A), chosen because of building block availability. Despite having a degree of molecular complexity, MAT-POS-b3e365b9-1 is obtained through a one-step amide coupling (Fig. 2A). We then further explored the P2 pocket with a library chemistry strategy in mind. Thus, guided by free-energy calculations (Fig. 2C), we first substituted the chromane for a tetrahydroisoquinoline to introduce a functionalizable handle (MAT-POS-3ccb8ef6-1; Fig. 5A), which maintained potency. Finally, we constructed a focused library realized through sulphonamide Schotten-Baumann coupling (fig. S14), furnishing an increase in both enzymatic inhibition and cellular antiviral efficacy. This work led to a potent antiviral chemical series (Fig. 5A) with a favorable safety profile, low brain penetrance (fig. S15 and data S5), and improved oral bioavailability but moderate in vitro–in vivo correlation in clearance (fig. S16 and data S5; all measured cellular antiviral data are available in data S6).

As an example for the aminopyridine lead series, we discuss antiviral efficacy, absorption, distribution, metabolism, and excretion (ADME) and pharmacokinetic (PK) characteristics of compound MAT-POS-e194df51-1. MAT-POS-e194df51-1 was profiled in SARS-CoV-2 antiviral assays across multiple cell lines, exhibiting a median effective concentration (EC_{50}) of 64 nM in A549-ACE2-TMPRSS2 cells and 126 nM in HeLa-ACE2 cells without measurable cytotoxicity (Fig. 5B). This is in line with overall cellular efficacy for the chemical series: Of 150 compounds with enzyme assay IC_{50} < 500 nM assessed in A549-ACE2-TMPRSS2 cellular cytophatic effect (CPE) assays, 15 compounds showed lower EC_{50} values compared with the internal control nirmatrelvir that was measured at an EC_{50} of 218 nM in this assay (Fig. 5C). Similarly, good antiviral activity was measured across "crowdsourced" antiviral assays across different laboratories and cell lines, including assays performed with and without p-gp inhibitors and using nirmatrelvir as an internal control (Fig. 5D). We also observed good cross-reactivity of our lead compound MAT-POS-e194df51-1 against known SARS-CoV-2 variants Alpha, Beta, Delta, and Omicron (Fig. 5E). Closely related molecules PET-UNK-29afea89-2 and MAT-POS-932d1078-3 with EC_{50} values in HeLa-ACE2 CPE assays of 240 nM and 331 nM and with values of 657 nM and 2.57 μM in A549-ACE2-TMPRSS2 CPE assays, respectively (fig. S17, A and B), show a >100-fold reduction of intracellular viral RNA and infectious virus secretion into the apical compartment of human induced pluripotent stem cell (iPSC)–derived kidney organoids (fig. S16, D and E)—an accessible model for the human kidney, an organ that is infected in COVID-19 patients—as reported previously for earlier analogs of the same series (*52*). MAT-POS-e194df51-1 exhibits favorable properties required for an orally bioavailable inhibitor (Fig. 5, F and G). In addition, crystallographic studies reveal that the interaction pattern of MAT-POS-e194df51-1 with the Mpro binding site is distinct to approved Mpro inhibitors nirmatrelvir and ensitrelvir (S-217622) (fig. S18), potentially offering complementary resistance profiles and justifying further development.

Open science presents a viable route to early drug discovery

The results presented here reflect the success of an open science, patent-free antiviral discovery

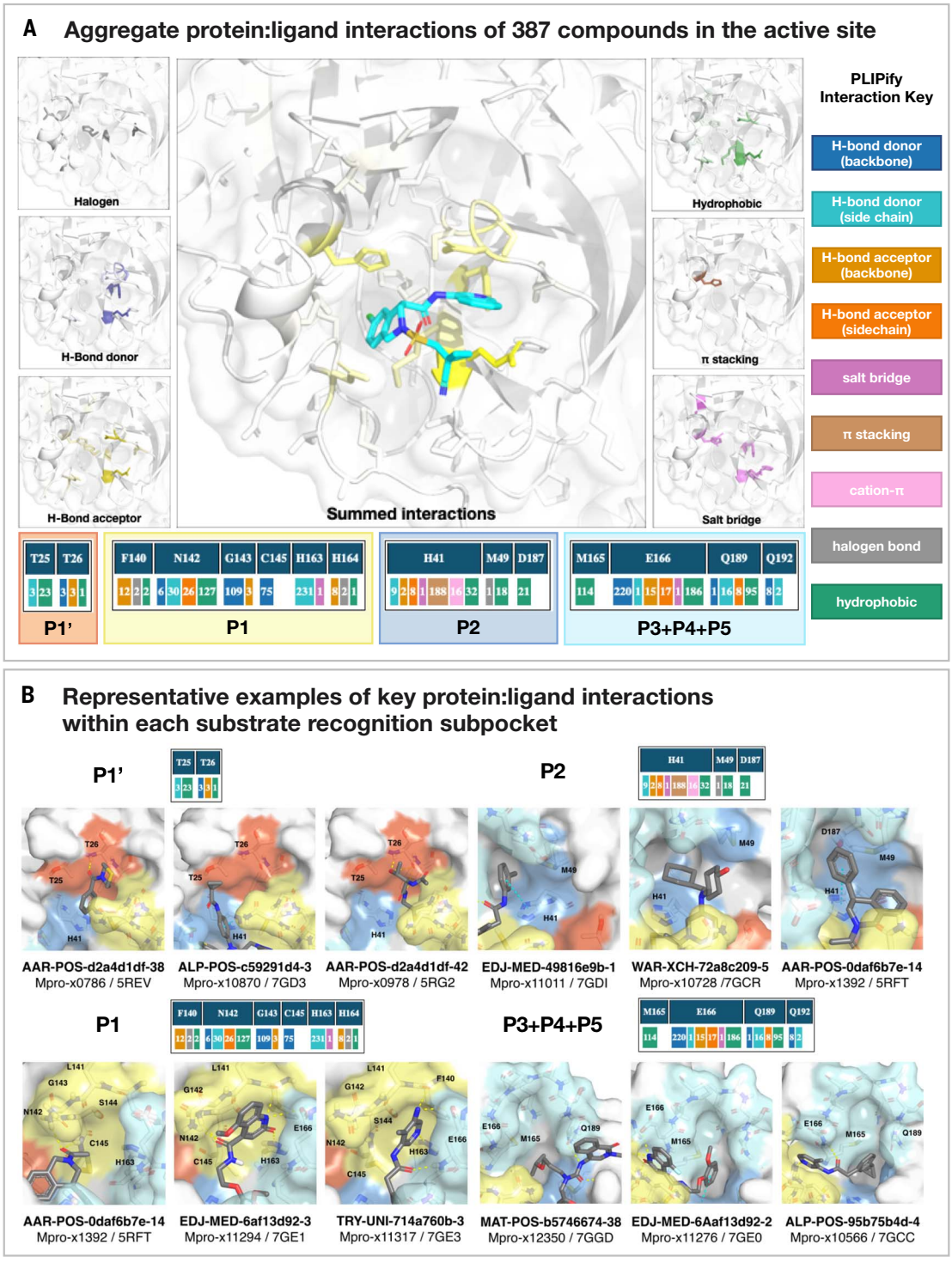


Fig. 3. Analysis of 367 complex crystal structures reveals hotspots for ligand engagement and a variety of ways to engage substrate recognition subpockets. (**A**) The five substrate recognition subpockets exhibit distinct preferences for intermolecular interactions. The figure highlights the locations of different types of interactions, with the shading indicating the frequency. The bottom row tallies the number of times that each interaction was seen in our structures for different residues. The interaction map was generated using PLIPify (Materials and methods) and summarizes the interactions witnessed across 367 complexes from the perspective of the protein, distinguishing between backbone (bb) and sidechain (sc) interactions (which might be more vulnerable to point mutations). (**B**) Representative examples of protein-ligand interactions engaging the P1', P1, P2, and P3-5 subpockets. Hydrogen bonds and π-stacking interactions are depicted as yellow and cyan dashed lines, respectively. The rows above each set of complexes tally the number of times that each interaction was seen with the specific residues within the subpockets. See data S4 for Protein Data Bank (PDB) IDs and crystallography statistics. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

program in rapidly developing a differentiated optimized lead in response to an emerging pandemic threat. As a result of the open science policy, a large number of collaborators (now the COVID Moonshot Consortium) were able to provide in-kind support, providing synthesis, assays, and in vitro and in vivo experiments. By making all data immediately available and all compounds purchasable from Enamine, we aim to accelerate research globally along parallel tracks following up on our initial work. As a notable example for the impact of open science, the Shionogi clinical candidate S-217622 [which has now received emergency approval in Japan as Xocova (ensitrelvir)] was identified in part on the basis of crystallographic data openly shared by the COVID Moonshot Consortium (*53*).

Despite our optimization and characterization efforts, considerable gaps from reporting a clinical candidate remain: The series requires further PK and pharmacodynamic (PD) optimization; in particular, it displays high clearance and low bioavailability. As it stands, it would likely not be able to achieve therapeutic exposure without a PK booster (such as ritonavir). To move forward, additional in-depth safety data are required as well as additional PK data from a second species to enable accurate human dose prediction. The COVID Moonshot and its lead series for COVID-19 have been adopted into the

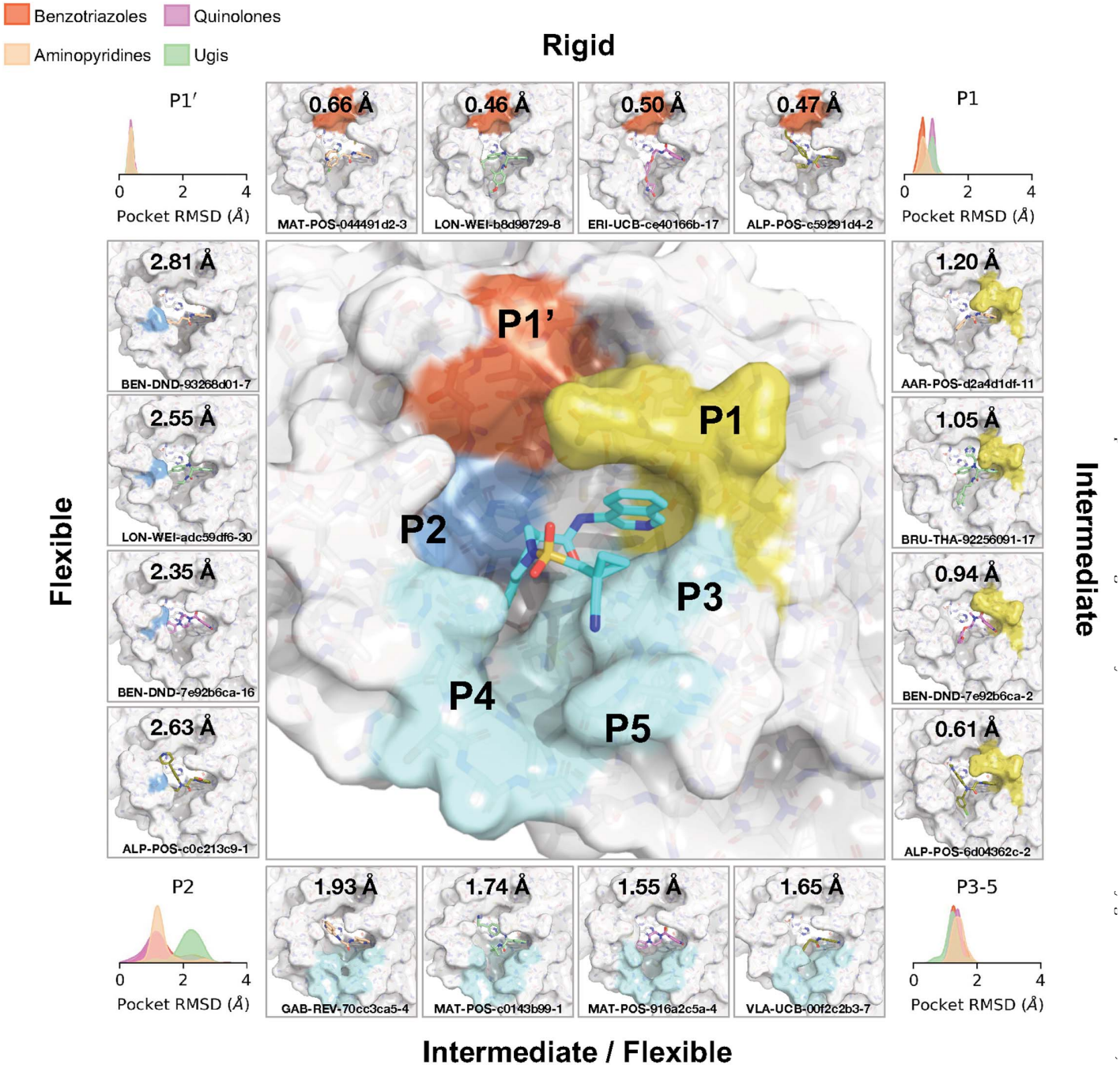


Fig. 4. Structural plasticity of the binding subpockets. The subpockets have different degrees of plasticity, which is also dependent on the chemical series (fig. S2). The corners of the figure show the distribution of sidechain root mean square deviation (RMSD) deviations from the structure of MAT-POS-e194df51-1 (middle panel; PDB: 7GAW). The boxes exemplify ligands that significantly deform the pockets.

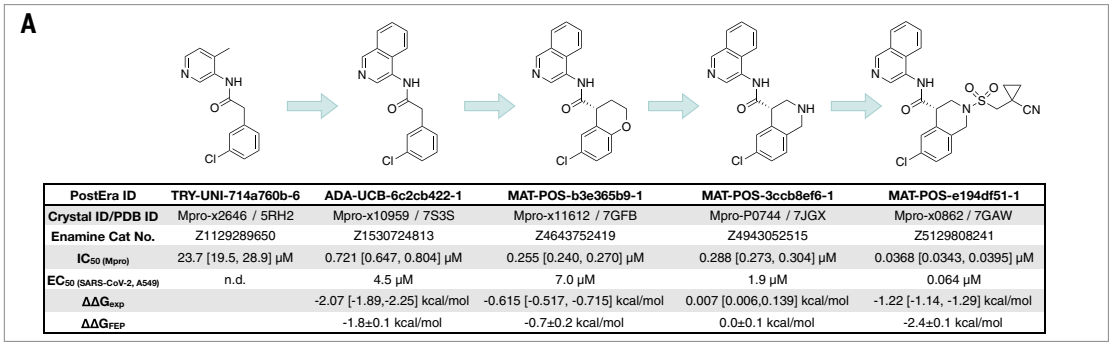
drug development portfolio of the Drugs for Neglected Diseases initiative (DND*i*) for further lead optimization and downstream preclinical development. This work is funded by a $10 million award from the Wellcome Trust through the World Health Organization (WHO) Access to COVID-19 Tools Accelerator (ACT-A) program, of which results will be reported upon filing Clinical Trials Authorization (CTA) (*54*). To reach phase 2 readiness, we expect a further $7.5 million will be required to process route development costs (*55*).

Open science efforts have transformed many areas of biosciences, with examples such as the Human Genome Project (*56*), the Structural Genomics Consortium (*57*), and the RAS Initiative (*58*). The COVID Moonshot provides an example of open science drug discovery leading to advances in infectious diseases drug discovery—a research area of grave public importance, but one that is chronically underfunded by the private sector (*59*).

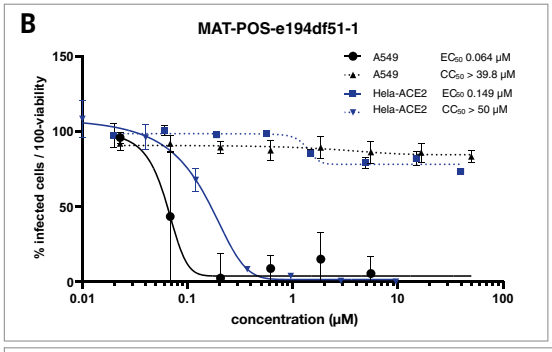
Materials and methods

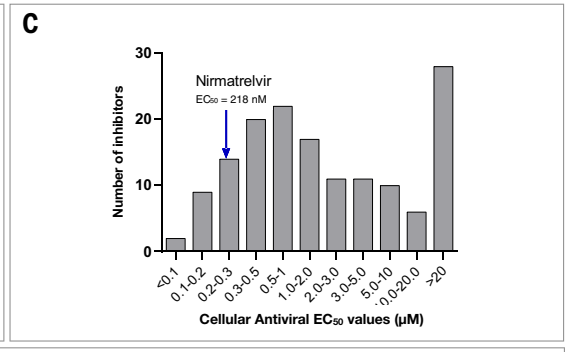
0. Compound registration and data flow process

All compound designs from the internal medicinal chemistry team, collaborators, and external submitters were captured through the online compound design submission platform (https://covid.postera.ai/covid) along with



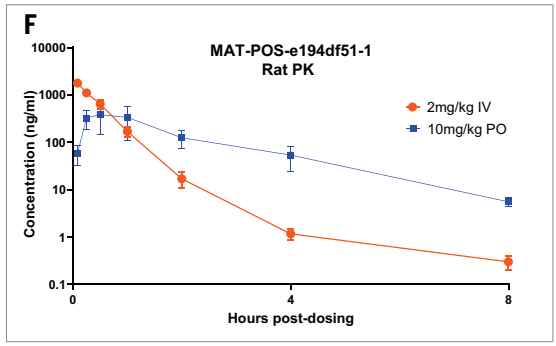
PostEra ID	TRY-UNI-714a760b-6	ADA-UCB-6c2cb422-1	MAT-POS-b3e365b9-1	MAT-POS-3ccb8ef6-1	MAT-POS-e194df51-1
Crystal ID/PDB ID	Mpro-x2646 / 5RH2	Mpro-x10959 / 7S3S	Mpro-x11612 / 7GFB	Mpro-P0744 / 7JGX	Mpro-x0862 / 7GAW
Enamine Cat No.	Z1129289650	Z1530724813	Z4643752419	Z4943052515	Z5129808241
$IC_{50\ (Mpro)}$	23.7 [19.5, 28.9] μM	0.721 [0.647, 0.804] μM	0.255 [0.240, 0.270] μM	0.288 [0.273, 0.304] μM	0.0368 [0.0343, 0.0395] μM
$EC_{50\ (SARS\text{-}CoV\text{-}2,\ A549)}$	n.d.	4.5 μM	7.0 μM	1.9 μM	0.064 μM
$\Delta\Delta G_{exp}$		-2.07 [-1.89,-2.25] kcal/mol	-0.615 [-0.517, -0.715] kcal/mol	0.007 [0.006,0.139] kcal/mol	-1.22 [-1.14, -1.29] kcal/mol
$\Delta\Delta G_{FEP}$		-1.8±0.1 kcal/mol	-0.7±0.2 kcal/mol	0.0±0.1 kcal/mol	-2.4±0.1 kcal/mol





Compound	A549 (+ P-gp inhibitor) CPE			Hela-ACE2 (+ P-gp inhibitor) CPE			Calu-3 (no P-gp inhibitor) CPE			VeroE6 (no P-gp inhibitor) CPE	
	EC_{50} (μM)	EC_{90} (μM)	CC_{50} (μM)	EC_{50} (μM)	EC_{90} (μM)	CC_{50} (μM)	EC_{50} (μM)	EC_{90} (μM)	CC_{50} (μM)	EC_{50} (μM)	EC_{90} (μM)
Nirmatrelvir	0.218	0.336	> 50	0.0604	0.12	> 39.8	2.01	6.08	> 100	2.71	3.71
MAT-POS-e194df51-1	0.0638	0.126	> 50	0.149	0.365	> 39.8	1.15	4.68	> 100	n.d.	n.d.
MAT-POS-24589f88-3	1.65	21.2	> 50	0.428	1.4	> 39.8	1.02	1.31	> 100	0.213	0.391
MAT-POS-3ccb8ef6-1	1.87	10.2	> 50	n.d.	n.d.	n.d.	2.57	6.96	> 100	0.513	2.16
MAT-POS-b3e365b9-1	6.97	> 50	> 50	0.481	1.29	> 39.8	1.06	3.43	> 100	1.57	3.48
ADA-UCB-6c2cb422-1	4.49	7.06	> 50	n.d.	n.d.	n.d.	> 10	n.d.	> 100	3.24	6.41
TRY-UNI-714a760b-6	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.

Virus	MAT-POS-e194df51-1		Nirmatrelvir	
	EC_{50} (μM)	CC_{50} (μM)	EC_{50} (μM)	CC_{50} (μM)
Alpha variant (B.1.1.7)	0.38	> 20	0.12	> 10
Beta variant (B.1.351)	1.48	> 20	0.21	> 10
Delta variant (B.1.617.2)	1.52	> 20	0.21	> 10
Omicron variant (B.1.529	0.29	> 20	0.07	> 10
MA-SARS-CoV-2/WA1	0.43	> 20	0.14	> 10



MAT-POS-e194df51-1

Antiviral Efficacy	
MPro IC_{50} (μM)	0.037
A549 EC_{50} (μM)	0.064

In Vitro ADME				
LogD (measured)	2.5			
MDCK-LE $P_{app\ A\rightarrow B}$ (10^{-6} cm/s)	21.8			
	Rat	Dog	Mining	Human
Liver microsomes CL (μL/min/mg)	604	164	542	152
Liver microsomes $t_{1/2}$ (min)	2.4	8.5	2.6	9.1
Hepatocyte CL (μL/min/10^6 cells)	67.6	61.4	65.9	10.3
Hepatocyte $t_{1/2}$ (min)	10.3	11.3	10.5	67.5
PPB free fraction (%)	5.4	n.d.	n.d.	10.1

Safety / Drug-Drug Interactions	
CYP Inhibition (μM) 2C9 /2D6 / 3A4	25 / 9.4 / 10.3
PXR risk	Low
hERG (μM)	> 30

In Vivo Pharmacokinetics	
Rat IV V_{ss} (L/kg)	1.05
Rat IV CL (mL/min/kg)	34.8
Rat $t_{1/2}$ IV/PO (h)	0.448 / 1.4
Rat Bioavailability (%)	18

Fig. 5. Iterative medicinal chemistry furnished an orally bioavailable inhibitor. (**A**) Summary of key medicinal chemistry milestones in developing the initial crowd-sourced lead compound into a potent antiviral. X-ray structures for each milestone compound are available via Fragalysis, and each compound can be obtained from Enamine using the corresponding catalog numbers. Retrospective alchemical free-energy calculation predictions for each transformation ($\Delta\Delta G_{FEP}$) are shown for each step between milestones, along with the corresponding experimental free-energy difference ($\Delta\Delta G_{exp}$) derived from biochemical activities. As positive control, under our assay condition, nirmatrelvir has an IC_{50} of 2.6 nM. (**B**) Antiviral activity of MAT-POS-e194df51-1 cellular antiviral assays, with an EC_{50} of 64 nM in A549-ACE2-TMPRSS2 cells assessing CPE (black; plotted as 100 − percent viability) and 126 nM in HeLa-ACE2 assays (blue; plotted as percentage of infected cells). Both assays were performed with P-gp inhibitors. (**C**) Histogram comparing antiviral efficacy of all COVID Moonshot compounds measured to date in an A549-ACE2-TMPRSS2 CPE cellular antiviral assay. (**D**) Detailed cellular antiviral assessment of key compounds composing the synthetic strategy (A) across different cell lines and assay techniques, with and without p-gp inhibitors, demonstrating efficacy of MAT-POS-e194df51-1 in various setups and laboratories. (**E**) MAT-POS-e194df51-1 shows good cross-reactivity against known circulating variants of SARS-CoV-2 in antiviral cellular assays in a CPE assay in HeLa-ACE2 cells. (**F**) PK profile of MAT-POS-e194df51-1 in rats with a 2 mg/kg intravenous and 10 mg/kg oral dosing with good oral availability. (**G**) ADME characteristics of MAT-POS-e194df51-1 demonstrate a favorable safety profile, indicating translational potential of the lead series.

submitter identity, institution, design rationale, and any inspiration fragments. A forum thread was created to discuss these designs and attached to the compound design. Each submitted batch of related designs received a unique ID including the first three letters of the submitter name and submitter institution, and each compound design submitted received a unique ID ("PostEra ID") that appended a unique molecule sequence ID within the submission batch ID. Internally, compound designs, synthesized compounds, and compounds with experimental data were tracked with corresponding records in a CDD Vault (Collaborative Drug Discovery Inc.).

Stereochemistry

Although the design platform enabled submitters to register compounds with specific

defined or uncertain stereochemistry, compounds were initially synthesized and biochemically assayed as racemates, and if active, chirally separated compounds were registered and assayed separately. Because the absolute stereochemical identity of enantiopure compounds was unknown at time of receipt, assay data were attached to compound records with specified relative stereochemistry, rather than absolute stereochemistry. For compounds where sufficient data were available from a variety of sources to propose the absolute stereochemistry (e.g., x-ray data for the compound or a close analog), the "suspected_SMILES" record was updated along with an articulated rationale in the "why_suspected_SMILES" field. As a result, caution must be exercised when using data for enantiopure compounds for downstream uses (e.g., whole-dataset machine learning) without verifying whether the absolute stereochemistry is known with confidence.

Submission analysis

The submitter names were standardized by removing affiliations and expansion of first name abbreviations, the submissions by two users who submitted large batches of compounds in an automated way in contravention of the goal of the project were removed. The word cloud was generated by filtering against 1000 most-common words and removing grammatical inflections and generating an image with an online word cloud generator. The classification of the methodology was done by presence of keywords determined by a simple keyword classifier with manually determined words (circa 100 training, 100 test) wherein "dock," "seesar," "vina," "autodock," "screen," "drug-hunter" were typical of docking, whereas "by-eye," "merg[ing]," "link[ing]," "coupl[ing]" were typical of hypothesis driven methods. A large fraction could not be accurately classified due to paucity of information. SAScore was calculated with Postera Manifold under the retrosynthesis route.

1. Experimental methods

1.1 Protease activity assays

1.1.1 Fluorescence Mpro inhibition assay

Compounds were seeded into assay-ready plates (Greiner 384 low volume, cat. no. 784900) using an Echo 555 acoustic dispenser, and dimethyl sulfoxide (DMSO) was back-filled for a uniform concentration in assay plates (DMSO concentration maximum 1%) Screening assays were performed in duplicate at 20 μM and 50 μM. Hits of greater than 50% inhibition at 50 μM were confirmed by dose response assays. Dose response assays were performed in 12-point dilutions of twofold, typically beginning at 100 μM. Highly active compounds were repeated in a similar fashion at lower concentrations beginning at 10 μM or 1 μM. Reagents for Mpro assay were dispensed into the assay plate in 10 μl volumes for a final volume of 20 μl.

Final reaction concentrations were 20 mM HEPES pH 7.3, 1.0 mM TCEP, 50 mM NaCl, 0.01% Tween-20, 10% glycerol, 5 nM Mpro, 375 nM fluorogenic peptide substrate ([5-FAM]-AVLQSGFR-[Lys(Dabcyl)]-K-amide). Mpro was pre-incubated for 15 min at room temperature with compound before addition of substrate and a further 30-min incubation. Protease reaction was measured in a BMG Pherastar FS with a 480/520 excitation/emission filter set. Raw data were mapped and normalized to high (Protease with DMSO) and low (No Protease) controls using Genedata Screener software. Normalized data were then uploaded to CDD Vault (Collaborative Drug Discovery). Dose response curves were generated for IC_{50} using nonlinear regression with the Levenberg–Marquardt algorithm with minimum inhibition = 0% and maximum inhibition = 100%.

The assay was calibrated at different enzyme concentrations to confirm linearity and response of protease activity, as well as optimization of buffer components for most stable and reproducible assay conditions. Substrate concentration was chosen after titration to minimize saturation of signal in the plate reader while obtaining a satisfactory and robust dynamic range of typically five- to sixfold over control without enzyme. We used low substrate concentrations of the bright FRET peptide to avoid "inner filter effect" (*60*) and to bias toward detection of competitive inhibitors (*61*). As positive control, under our assay condition, nirmatrelvir has IC_{50} of 2.6 nM.

1.1.2 RapidFire Mpro inhibition assay

The assay was performed according to the published procedure (*62*). Briefly, compounds were seeded into assay-ready plates (Greiner 384PP, cat. no. 781280) using an ECHO 650T dispenser and DMSO was back-filled for a uniform concentration in assay plates (DMSO concentration < 1%, final volume = 500 nl.). A 15 μM enzyme stock solution was prepared in 20 mM HEPES, pH 7.5 and 300 mM NaCl, and subsequently diluted to a working solution of 300 nM Mpro in assay buffer (20 mM HEPES, pH 7.5 and 50 mM NaCl) before the addition of 25 μl to each well using a Multidrop Combi (Thermo Scientific). After a quick centrifugation step (1000 rpm, 15 s) the plate was incubated for 15 min at room temperature. The reaction is initiated with the addition of 25 μl of 4 μM 11-nucleotide oligomer (TSAVLQSGFRK-NH2, initially custom synthesized by the Schofield group, then by GLBiochem, used until March 2021), or 10 μM 37-nucleotide oligomer (ALNDFSNSGS-DVLYQPPQTSITSAVLQSGFRKMAFPS-NH2, GLBiochem, used after March 2021), dissolved in assay buffer. After centrifugation (1000 rpm, 14 s) the reaction is incubated for 10 min (11-nucleotide oligomer) or 5 min (37-nucleotide oligomer) at room temperature before quenching with 10% formic acid. The reactions are analyzed with MS using RapidFire (RF) 365 high-throughput sampling robot (Agilent) connected to an iFunnel Agilent 6550 accurate mass quadrupole time-of-flight (Q-TOF) mass spectrometer using electrospray. All compounds are triaged by testing the percentage inhibition at 5 and 50 μM final concentration. Dose response curves uses an 11-point range of 100 to 0.0017 μM inhibitor concentrations. RF integrator software (Agilent) was used to extract the charged states from the total ion chromatogram data followed by peak integration. For the 11-nucleotide oligomer peptide the m/z (+1) charge states of both the substrate (1191.67 Da) and cleaved N-terminal product TSAVLQ (617.34 Da) were used and the 37-nucleotide oligomer peptide the m/z (+2) charge states of the substrate (3960.94 Da) and m/z (+1) of the cleaved C-terminal product SGFRKMAFPS (1125.57 Da). Percentage conversion [(product peak integral)/(product peak integral + substrate peak integral) × 100] and percentage inhibitions were calculated and normalized against DMSO control with deduction of any background signal in Microsoft Excel. IC_{50} values were calculated using Levenberg–Marquardt algorithm used to fit a restrained Hill equation to the dose-response data with both GraphPad PRISM and CDD.

1.2 High-throughput x-ray crystallography

Purified protein (*20*) at 24 mg/ml in 20 mM HEPES pH 7.5, 50 mM NaCl buffer was diluted to 12 mg/ml with 20 mM HEPES pH 7.5, 50 mM NaCl before performing crystallization using the sitting-drop vapor diffusion method with a reservoir solution containing 11% PEG 4 K, 5% DMSO, 0.1 M MES pH 6.5. Crystals of Mpro in the monoclinic crystal form (C2), with a single monomer in the asymmetric unit, were grown with drop ratios of 0.15 μl protein, 0.3 μl reservoir solution, and 0.05 μl seeds prepared from previously produced crystals of the same crystal form (*20*). Crystals in the orthorhombic crystal form ($P2_12_12_1$), with the Mpro dimer present in the asymmetric unit, were grown with drop ratios of 0.15 μl protein, 0.15 μl reservoir solution, and 0.05 μl seeds prepared from crystals of an immature Mpro mutant in the same crystal form (*63*).

Compounds were soaked into crystals by adding compound stock solutions directly to the crystallization drops using an ECHO liquid handler. In brief, 40 to 90 nl of DMSO solutions (between 20 and 100 mM) were transferred directly to crystallization drops using giving a final compound concentration of 2 to 20 mM and DMSO concentration of 10 to 20%. Drops were incubated at room temperature for ~1 to 3 hours before mounting and flash cooling in liquid nitrogen without the addition of further cryoprotectant.

Data were collected at Diamond Light Source on the beamline I04-1 at 100 K and processed

with the fully automated pipelines at Diamond (*64*–*66*), which include XDS (*67*), xia2 (*68*), autoPROC (*69*), and DIALS (*64*). Further analysis was performed using XChemExplorer (*70*) with electron density maps generated using DIMPLE (http://ccp4.github.io/dimple/). Ligand-binding events were identified using PanDDA (*71*) (https://github.com/ConorFWild/pandda), and ligands were manually modeled into PanDDA-calculated event maps or electron density maps using Coot (*72*). Ligand restraints were calculated with ACEDRG (*73*) or GRADE [grade v. 1.2.19 (Global Phasing Ltd., Cambridge, UK, 2010)] and structures refined with Buster [Buster v. 2.10.13 (Cambridge, UK, 2017)]. Models and quality annotations were reviewed using XChemReview (*74*), Buster-Report [Buster v. 2.10.13 (Cambridge, UK, 2017)] and Mogul (*75*, *76*).

Coordinates, structure factors and PanDDA event maps for all datasets are available on Fragalysis (https://fragalysis.diamond.ac.uk/viewer/react/preview/target/Mpro).

1.3 Viral screening assays

A variety of antiviral replication assays were performed in collaborating laboratories, including cytopathic effect (CPE) inhibition assays at the IIBR, Israel, and Katholieke Universiteit Leuven; quantitative reverse-transcription polymerase chain reaction (RT-qPCR) for viral RNA at Radboud University Medical Center, Netherlands; immunofluorescence assays at University of Nebraska Medical Center, USA; and plaque assays and focus-forming unit (FFU) assays at University of Oxford, UK.

1.3.1 Antiviral cytopathic effect assay, VeroE6 (IIBR, Ness-Ziona, Israel)

SARS-CoV-2 (GISAID accession EPI_ISL_406862) was kindly provided by Bundeswehr Institute of Microbiology, Munich, Germany. Virus stocks were propagated (4 passages) and tittered on Vero E6 cells. Handling and working with SARS-CoV-2 virus was conducted in a BSL3 facility in accordance with the biosafety guidelines of the Israel Institute for Biological Research (IIBR). Vero E6 were plated in 96-well plates and treated with compounds in medium containing 2% fetal bovine serum (FBS). The assay plates containing compound dilutions and cells were incubated for 1 hour at 37°C before adding multiplicity of infection (MOI) 0.01 of viruses. Viruses were added to the entire plate, including virus control wells that did not contain test compound and Remdesivir drug used as positive control. After 72 hours incubation, viral CPE inhibition assay was measured with XTT reagent. Three replicate plates were used.

1.3.2 Antiviral immunofluorescence assay, VeroE6 (Pathology and Microbiology, University of Nebraska Medical Center, USA, St Patrick Reid)

Vero E6 cells were pretreated with 20 uM of the Moonshot compounds for around 2 hours. Cells were then infected with SARS-CoV-2 at a MOI of 0.1 for 24 hours. Virus infection was terminated by 4% paraformaldehyde (PFA) fixation. Cells were stained using a Rabbit SARS-CoV-2 antibody (Sino Biological 40150-R007) as a primary antibody, and Alexa-488, Hoechst and Cell Mask (Thermo Fisher) as a secondary antibody. Images were collected on the Operetta system imaging system and analyzed using the Harmony software.

1.3.3 Antiviral FFU assay, Calu-3 (University of Oxford, UK)

Cell culture

The African green monkey Vero E6 cell line (ATCC CRL-1586) was cultured in Dulbecco's modified Eagle medium (DMEM) with Glutamax supplemented with 100 μg/ml streptomycin, 100 U/ml penicillin, and 10% heat-inactivated fetal calf serum (FCS). The human lung cancer cell line Calu-3 (Anderson Ryan, Department of Oncology, Medical Science Division, University of Oxford) was cultured in a 1:1 mixture of DMEM with Glutamax and Ham's F-12 medium supplemented with 100 μg/ml streptomycin, 100 U/ml penicillin, and 10% heat-inactivated FCS. All cells were maintained as mycoplasma free, with regular verifications by polymerase chain reaction (PCR).

Virus propagation

SARS-CoV-2 England/2/2020 was provided at passage 1 from Public Health England, Collindale. Passage 2 submaster and passage 3 working stocks were produced by infecting Vero E6 cells at a MOI of 0.01 in virus propagation medium (DMEM with Glutamax supplemented with 2% FCS) and incubating until CPE was visible. The cell supernatant was then centrifuged at 500 g for 5 min, aliquoted and stored at −80°C. The titer of viral stocks was determined by plaque assay. All subsequent assays were performed using a passage 3 stock.

Cell viability

Cell viability was measured using the CellTiter 96 R AQueous One Solution Cell Proliferation MTA [3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H - 15 tetrazolium, inner salt] Assay (Promega) according to the manufacturer's instruction after treatment with compound. Briefly, Calu 3 cells were treated with compounds in quadruplicate for 3 days. Wells with 200 μl growth medium with and without cells were included as controls in quadruplicate. After the incubation, 100 μl of growth medium was removed and 20 μl of MTS reagent was added to the remaining medium in each well. After a further 1- to 2-hour incubation, the absorbance at 490 nm was measured on a Molecular Devices SpectraMax M5 microplate reader.

Antiviral assays

For FFU assays, a SARS-CoV-2 Microneutralization assay from the W. James laboratory (Dunn School of Pathology, University of Oxford) was adapted for use as a FFU assay. Briefly, 3 half log dilutions of each supernatant to be analyzed were prepared in virus propagation medium. 20 μl of each dilution was inoculated into wells of a 96-well plate in quadruplicate followed by 100 μl Vero E6 cells at 4.5×10^5 cells/ml in virus propagation medium. The plates were incubated for 2 hours before the addition of 100 μl of 1.8% CMC overlay, and then incubated for a further 24 hours. After 24 hours the overlay was carefully removed and the cells washed once with PBS before fixing with 50 μl of 4% PFA, after 30 min the PFA was removed and replaced with 100 μl of 1% ethanolamine in PBS. The cells were permeabilized by replacing the ethanolamine with 2% Triton X100 in PBS and incubating at 37°C for 30 min. The plates were then washed three times with wash buffer (0.1% Tween 20 in PBS) inverted and gently tapped onto tissue to dry before the addition of 50 μl of EY2A anti-N human monoclonal antibody (mAb) [Arthur Huang (Taiwan)/Alain Townsend (Weatherall Institute of Molecular Medicine, University of Oxford)] at 10 pmol in wash buffer. The plates were rocked at room temperature for 1 hour, washed and incubated with 100 μl of secondary antibody anti-human immunoglobulin G (IgG) (Fc-specific)-peroxidase-conjugate produced in Goat diluted 1:5000 at room temperature for 1 hour. 50 μl of TrueBlue peroxidase substrate was added to the wells and incubated at RT for 10 min on the rocker, after 10 min the substrate was removed, and the plates washed with ddH2O for 10 min. The water was removed and the plates allowed to air dry. The foci were then counted using an ELISPOT classic reader system (AID GmbH).

1.3.4 Antiviral qPCR assay, Vero E6 and kidney organoids (Radboud University Medical Center, Nijmegen, Netherlands)

Cell culture

African green monkey Vero E6 kidney cells (ATCC CRL-1586) and Vero FM kidney cells (ATCC CCL-81) were cultured in DMEM with 4.5 g/L glucose and L-glutamine (Gibco), supplemented with 10% FCS (Sigma Aldrich), 100 μg/ml streptomycin and 100 U/ml penicillin (Gibco). Cells were maintained at 37°C with 5% CO_2. Human iPSC-derived kidney organoids were prepared as previously described (*52*).

Virus propagation

SARS-CoV-2 (isolate BetaCoV/Munich/BavPat1/2020) was kindly provided by C. Drosten (Charité-Universitätsmedizin Berlin, Institute of Virology, Berlin, Germany) and was initially cultured in Vero E6 cells up to three passages in the laboratory of Bart Haagmans (Viroscience Department, Erasmus Medical Center, Rotterdam, Netherlands). Vero FM cells were infected with passage 3 stock at an MOI of 0.01 in infection medium (DMEM containing L-glutamine, 2%

FCS, 20 mM HEPES buffer, 100 μg/ml streptomycin and 100 U/ml penicillin). Cell culture supernatant containing virus was harvested at 48 hours postinfection (hpi), centrifuged to remove cellular debris, filtered using a 0.2 μm syringe filter (Whatman), and stored in 100 μl aliquots at −80°C.

Virus titration

Vero E6 cells were seeded in 12-well plates at a density of 500,000 cells per well. Cell culture medium was discarded at 24 hours postseeding, cells were washed twice with PBS and infected with 10-fold dilutions of the virus stock in unsupplemented DMEM. At 1 hpi, cells were washed with PBS and replaced with overlay medium, consisting of minimum essential medium (Gibco), 2% FCS, 20 mM HEPES buffer, 100 μg/ml streptomycin, 100 U/ml penicillin, and 0.75% carboxymethyl cellulose (Sigma Aldrich). At 72 hpi, the overlay medium was discarded, cells were washed with PBS and stained with 0.25% crystal violet solution containing 4% formaldehyde for 30 min. Afterward, staining solution was discarded and plates were washed with PBS, dried and plaques were counted.

Antiviral assay

Vero E6 cells were seeded onto 24-well plates at a density of 150,000 cells per well. At 24 hours postseeding, cell culture medium was discarded, cells were washed twice with PBS and infected with SARS-CoV-2 at an MOI of 0.01 in the presence of six concentrations of the inhibitors (25 μM to 0.06 μM). At 1 hpi, the inoculum was discarded, cells were washed with PBS, and infection medium containing the same concentration of the inhibitors was added to the wells. SARS-CoV-2 infection in the presence of 0.1% DMSO was used as a negative control. At 24 hpi, 100 μl of the cell culture supernatant was added to RNA-Solv reagent (Omega Bio-Tek) and RNA was isolated and precipitated in the presence of glycogen according to manufacturer's instructions. TaqMan Reverse Transcription reagent and random hexamers (Applied Biosystems) were used for cDNA synthesis. Semi-quantitative real-time PCR was performed using GoTaq qPCR (Promega) BRYT Green Dye-based kit using primers targeting the SARS-CoV-2 E protein gene (*77*) (forward primer, 5′-ACAGGTACGTTAATAGTTAATAGCGT-3′; reverse primer, 5′-ACAGGTACGTTAATAGTTAATAGCGT-3′). A standard curve of a plasmid containing the E gene qPCR amplicon was used to convert Ct values relative genome copy numbers. For viability assays, Vero E6 cells were seeded in 96-well white-bottom culture plates (Perkin Elmer) at a density of 30,000 cells per well. At 24 hours postseeding, cells were treated with the same concentrations of compounds as used for the antiviral assay. Cells treated with 0.1% DMSO were used as a negative control. At 24 hours post-treatment, cell viability was assessed using the Cell Titer Glo 2.0 kit (Promega) using the Victor Multilabel Plate Reader (Perkin Elmer) to measure luminescence signal.

Antiviral assays in organoids

Human iPSC-derived kidney organoids cultured in transwell filters (Corning) were infected with SARS-CoV-2 in the presence of 1 and 10 μM of MAT-POS-932d1078-3, PET-UNK-29afea89-2 or 0.1% DMSO using an MOI of 1.0 in Essential 6 medium (Gibco) at 37°C and 5% CO_2, exposing the cells both basolaterally and apically to the inoculum. After 24 hours, medium containing the inoculum was removed and fresh essential 6 medium containing the same concentration of inhibitor was added to the basolateral compartment and cells were cultured for an additional 24 hours. At 48 hpi, organoids were washed in PBS, and the apical surface was exposed to Essential 6 medium for 10 min at 37°C, which was collected and used for viral titration. Individual organoids were harvested for RNA isolation using the PureLink RNA mini kit (Thermo Fisher) according to manufacturer's instructions. Viral RNA copies were analyzed by RT-qPCR on the SARS-CoV E gene, as described previously (*78*).

1.3.5 High-content SARS-CoV-2 antiviral screening assay, HeLa-ACE2 (Takeda via Calibr/TSRI)

SARS-CoV-2/HeLa-ACE2 high-content screening assay

Compounds are acoustically transferred into 384-well μclear-bottom plates (Greiner, part no. 781090-2B) and HeLa-ACE2 cells are seeded in the plates in 2% FBS at a density of 1.0×10^3 cells per well. Plated cells are transported to the BSL3 facility where SARS-CoV-2 (strain USA-WA1/2020 propagated in Vero E6 cells) diluted in assay media is added to achieve ~30 to 50% infected cells. Plates are incubated for 24 hours at 34°C 5% CO_2, and then fixed with 8% formaldehyde. Fixed cells are stained with human polyclonal sera as the primary antibody, goat anti-human H+L conjugated Alexa 488 (Thermo Fisher Scientific A11013) as the secondary antibody, and antifade 4′,6-diamidino-2-phenylindole (DAPI) (Thermo Fisher Scientific D1306) to stain DNA, with PBS 0.05% Tween 20 washes in between fixation and subsequent primary and secondary antibody staining. Plates are imaged using the ImageXpress Micro Confocal High-Content Imaging System (Molecular Devices) with a 10× objective, with four fields imaged per well. Images are analyzed using the Multi-Wavelength Cell Scoring Application Module (MetaXpress), with DAPI staining identifying the host-cell nuclei (the total number of cells in the images) and the SARS-CoV-2 immunofluorescence signal leading to identification of infected cells.

Uninfected host cell cytotoxicity counter screen

Compounds are acoustically transferred into 1536-well plates (Corning no. 9006BC). HeLa-ACE2 cells are maintained as described for the infection assay and seeded in the assay-ready plates at 400 cells per well in DMEM with 2% FBS. Plates are incubated for 24 hours at 37°C 5% CO_2. To assess cell viability, 2 ml of 50% CellTiter Glo (Promega no. G7573) diluted in water is added to the cells and luminescence measured on an EnVision Plate Reader (Perkin Elmer).

Data analysis

Primary in vitro screen and the host cell cytotoxicity counter screen data are uploaded to Genedata Screener, Version 16.0. Data are normalized to neutral (DMSO) minus inhibitor controls (2.5 μM remdesivir for antiviral effect and 10 μM puromycin for infected host cell toxicity). For the uninfected host cell cytotoxicity counter screen 40 μM puromycin (Sigma) is used as the positive control. For dose response experiments compounds are tested in technical triplicates on different assay plates and dose curves are fitted with the four parameter Hill Equation.

1.3.6 Cytopathic effect assay, hACE2-TMPRSS2 cells (Katholieke Universiteit Leuven)

Virus isolation and virus stocks

All virus-related work was conducted in the high-containment BSL3 facilities of the KU Leuven Rega Institute (3CAPS) under licenses AMV 30112018 SBB 219 2018 0892 and AMV 23102017 SBB 219 2017 0589 according to institutional guidelines. The SARS-CoV-2 strain used for this study was the Alpha variant of Concern (derived from hCoV-19/Belgium/rega-12211513/2020; EPI_ISL_791333, 2020-12-21). Virus sample was originally isolated in-house from nasopharyngeal swabs taken from travelers returning to Belgium (baseline surveillance) and were subjected to sequencing on a MinION platform (Oxford Nanopore) directly from the nasopharyngeal swabs. Virus stocks were then grown on Vero E6 cells in (DMEM 2% FBS medium) and passaged one time on A549-ACE2-TMPRSS2 cells. Median tissue culture infectious doses (TCID50) was defined by endpoint titration.

A549-ACE2-TMPRSS2 assay

A549-Dual hACE2-TMPRSS2 cells obtained by Invitrogen (cat. no. a549d-cov2r) were cultured in DMEM 10% FCS (Hyclone) supplemented with 10 μg/ml blasticidin (Invivogen, ant-bl-05), 100 μg/ml hygromycin (Invivogen, ant-hg-1), 0.5 μg/ml puromycin (Invivogen, ant-pr-1) and 100 μg/ml zeocin (Invivogen, ant-zn-05). For antiviral assay, cells were seeded in assay medium (DMEM 2%) at a density of 15,000 cells per well. One day after, compounds were serially diluted in assay medium (DMEM supplemented with 2% v/v FCS) and cells were infected with their respective SARS-CoV-2 strain at a MOI of ~0.003 TCID50/ml. On day 4 pi., differences in cell viability caused by virus-induced CPE or by

compound-specific side effects were analyzed using MTS as described previously (*79*). Cytotoxic effects caused by compound treatment alone were monitored in parallel plates containing mock-infected cells.

1.3.6 Immunofluorescence SARS-CoV-2 antiviral screening assay, HeLa-ACE2 (Mount Sinai)

Assessment of cross-reactivity against SARS-CoV-2 variant strains and cytotoxicity assays were performed as previously described (*80*). In brief, two thousand HeLa-ACE2 cells (BPS Bioscience) were seeded into 96-well plates in DMEM (10% FBS) and incubated for 24 hours at 37°C, 5% CO2. Two hours before infection, the medium was replaced with 100 µl of DMEM (2% FBS) containing the compound of interest at concentrations 50% greater than those indicated, including a DMSO control. Plates were then transferred into the BSL3 facility and 100 PFU (MOI = 0.025) was added in 50 µl of DMEM (2% FBS), bringing the final compound concentration to those indicated. Plates were then incubated for 48 hours at 37°C. After infection, supernatants were removed, and cells were fixed with 4% formaldehyde for 24 hours before being removed from the BSL3 facility. The cells were then immunostained for the viral N protein (an inhouse mAb 1C7, provided by Thomas Moran, thomas.moran@mssm.edu) with a DAPI counterstain. Infected cells (488 nm) and total cells (DAPI) were quantified using the Celigo (Nexcelcom) imaging cytometer. Infectivity was measured by the accumulation of viral N protein (fluorescence accumulation). Percent infection was quantified as [(infected cells/total cells) – background] × 100, and the DMSO control was then set to 100% infection for analysis. Data were fit using nonlinear regression and IC_{50} values for each experiment were determined using GraphPad Prism version 8.0.0 (San Diego, CA). Cytotoxicity was also performed using the MTT assay (Roche), according to the manufacturer's instructions. Cytotoxicity was performed in uninfected cells with same compound dilutions and concurrent with viral replication assay. All assays were performed in biologically independent triplicates.

2. Computational methods

2.1 Synthetic route planning

We use an approach based on the Molecular Transformer technology (*32*). Our algorithm uses natural language processing to predict the outcomes of chemical reactions and design retrosynthetic routes starting from commercially available building blocks. This proprietary platform is provided free of charge by PostEra Inc (https://postera.ai/). Additionally, Manifold (https://app.postera.ai/manifold/) was built by PostEra Inc. during the project to search the entire space of purchasable molecules, and automatically find the optimal building blocks.

2.2 Alchemical free-energy calculations

Large-scale alchemical free-energy calculations were conducted in batches ("Sprints") in which each set of calculations aimed to prioritize compounds that could be produced from a common synthetic intermediate using Enamine's extensive building block library, resulting in synthetic libraries of hundreds to tens of thousands. Virtual synthetic libraries were organized into a star map, where all transformations were made with respect to a single reference x-ray structure and compound with experimentally measured bioactivity. x-ray structures were prepared using the OpenEye Toolkit SpruceTK with manually controlled protonation states for the key His^{41}:Cys^{145} catalytic dyad (variously using zwitterionic or uncharged states) and His^{163} in P1 (which interacts with the 3-aminopyridine or isoquinoline nitrogen in our primary lead series). As the most relevant protonation states were uncertain, when computational resources afforded, calculations were carried out using multiple protonation state variants (His^{41}:Cys^{145} either neutral or zwitterionic; His^{163} neutral or protonated) and the most predictive model on available retrospective data for that scaffold selected for nominating prospective predictions for that batch. Initial poses of target compounds were generated via constrained conformer enumeration to identify minimally clashing poses using Omega (from the OpenEye Toolkit) using a strategy that closely follows an exercise described in a blog post by Pat Walters (http://practicalcheminformatics.blogspot.com/2020/03/building-on-fragments-from-diamondxchem_30.html). Alchemical free-energy calculations were then prepared using the open source perses relative alchemical free-energy toolkit (*40*) (https://github.com/choderalab/perses), and nonequilibrium switching alchemical free-energy calculations (*81*) were run on Folding@home using the OpenMM compute core (*43*). Nonequilibrium switching calculations used 1 ns nonequilibrium alchemical trajectories, where most calculations were performed with 1 fs time steps without constraints to hydrogen due to technical limitations that have been resolved in calculations using OpenMM 7.5.1 and later. We used the Open Force Field Initiative OpenFF "Parsley" small molecule force fields (*39*) (multiple generations between 1.1.1 and 1.3.1 were released and used as the project evolved) and the AMBER14SB protein force field (*82*) with recommended ion parameters (*83*, *84*), and TIP3P water (*85*). As many assayed compounds as possible were included in each batch of transformations to enable continual retrospective assessment and to leverage existing measured affinities in extrapolating predicted affinities. Analysis of free-energy calculations used the maximum likelihood estimator (*86*) to reconstruct the optimal predicted absolute free energy (and hence pIC_{50}) estimate from available experimental measurements. Calculations were analyzed using the fah-xchem dashboard (https://github.com/choderalab/fah-xchem) using the Bennett acceptance ratio (*87*, *88*) (https://threeplusone.com/pubs/gecthesis) and posted online in real time for the medicinal chemistry team to consult in making decisions about which compounds to prioritize.

We note that our primary aim was computing estimates of relative binding free energies for large alchemical transformations using abundant computing resources [which exceeded 1 exaFLOP/s (*38*)] rather than aggressive optimization of the cost/transformation. Batches of transformations used between 100 and 200 parallel 4 ns nonequilibrium cycles per transformation, selected based on the number of atoms modified in the transformation, resulting in 100 to 200 ns per transformation in aggregate. A Tesla V100 achieves ~200 ns/day for our solvated Mpro complex, meaning ~2 to 4 GPU-days per transformation was consumed on a V100 equivalent GPU. To give typical scales, Fig. 2C, panel 1, ran 6319 transformations of 140 cycles, resulting in ~3.5 ms of simulation time or ~424K GPU-hours; Fig. 2C, panel 2, ran 5077 transformations of ~200 cycles, resulting in ~4 ms simulation time, or ~480K GPU-hours; Fig. 2C, panel 3, ran 686 transformations of ~200 cycles, resulting in ~548 µs of simulation time, or ~66K GPU-hours.

Scripts for setting up and analyzing the perses alchemical free-energy calculations on Folding@home, as well as an index of computed datasets and dashboards are available at https://github.com/foldingathome/covid-moonshot

Code used for generating the COVID Moonshot alchemical free-energy calculation web dashboards is available here: https://github.com/choderalab/fah-xchem

Retrospective calculations for transformations in the main synthetic series shown in Fig. 5A were performed with an early release of perses 0.10.2 constructed as a simplified example that anyone can run to illustrate how these calculations work on standard GPU workstations, and use standard alchemical replica exchange protocols of 5 ns per replica (which just take a few hours on standard workstations, as opposed to the expensive nonequilibrium protocols used in the Sprints). Input scripts for this calculation are available in the perses distribution under "examples/moonshot-mainseries/" (https://github.com/choderalab/perses/tree/main/examples/moonshot-mainseries).

2.3 Structural flexibility and interactions analysis

Protein-ligand interactions are the driving forces for molecular recognition. In this work, the *PLIPify* repo (https://github.com/volkamerlab/plipify)

is used to detect shared interaction hotspots within the different Mpro structures. *PLIPify* is a python wrapper built on top of PLIP (*89*), a tool that enables automatic generation of protein-ligand interaction profiles for single complexes, to allow combining these profiles for multiple structures.

To generate the hotspots (depicted in Fig. 3A), the fragalysis data were downloaded (as of July 2021, https://fragalysis.diamond.ac.uk/api/targets/?format=json&title=Mpro). The respective pre-aligned complex structures were further investigated (found under data/{target}/aligned/{crystal_name}/{crystal_name}_bound.pdb). Only one chain per structure is kept, and the structures are protonated using Amber's *reduce* function. *PLIPify* is invoked, and structures are excluded from further analysis if they do not contain exactly one binding site (i.e., PLIP detects either zero or more than one binding sites), the sequence contains gaps ('-'), or the sequence length differs more than a standard deviation from the average length across all investigated structures.

This procedure resulted in a final set of 367 complex structures, used to generate the interaction fingerprints. Note for this study, only hbond-donor, hbond-acceptor, salt bridge, hydrophobic, pi-stacking, and halogen interactions are inspected. Additional code was added to *PLIPify* to split the hbond-donor and hbond-acceptor interactions into backbone and sidechain interactions (https://github.com/volkamerlab/plipify/pull/18). Interacting residues are only included if the summed interaction count per residue over all investigated structures is greater than five. Careful examination of examples of the interactions led us to filter out the S144 interactions from the final report as none of the interactions were convincing (24 hbond-don-bb, 168 hbond-don-sc, and 4 hbond-acc-sc interactions). The resulting structural depiction (Fig. 3A) were generated using pymol, and structure Mpro-P1788_0A_bound_chainA (protonated) is displayed (scripts available at https://github.com/volkamerlab/plipify/blob/master/projects/01/fragalysis.ipynb). Finally, structures containing compounds exhibiting some of the major interactions identified were used to generate the figures in Fig. 3B.

4. Chemical methods

4.1 HTC library synthesis

4.1.1 Chan-Lam reaction

The arylamine library was made by reacting the boronic acid (fig. S7D), under the optimized reaction conditions (1 eq. amine; 0.2 eq. CuI; 0.8 eq. DMAP; 2 eq. Hex_3N; DMSO; under air; RT; 2 days) with 296 amines (200 aromatic, 48 primary, and 48 secondary aliphatic amines; data S2). For library production, we used Echo LDV plates and an Echo 555 acoustic dispenser for liquid handling. After the allotted reaction time, plate copies were made after diluting the reaction mixture with 4.6 µl DMSO and transferring 1 µl of the obtained solution to a 384-well plate, for either biochemical assay or yield estimation.

4.1.2 Amide coupling

The amide library was made by reacting the carboxylic acid (fig. S8E) under the optimized reaction conditions (2 eq. amine; 2 eq. EDC; 2 eq. HOAt; 5 eq. DIPEA; DMSO; RT; 24 hours) with 300 amines (202 aromatics, 49 primary, and 49 secondary aliphatic amines; data S2). For library production, we used Echo LDV plates and an Echo 555 acoustic dispenser for liquid handling. Plate copies were made after diluting the reaction mixture with 4 µl DMSO. For yield estimation, 1 µl of the diluted library was transferred to an LC/MS-ready 384-well plate, followed by dilution with 20% ACN in water to the final volume of 50 µl. The desired product was identified in 60% of wells.

4.2 General compounds synthesis and characterization

All compounds were directly purchased from Enamine Inc., following Enamine's standard quality control (QC) for compound collections. In addition, in the supplementary chemistry section of the supplementary materials, we discuss the synthesis procedure, as well as liquid chromatography–mass spectrometry (LC-MS) and 1H nuclear magnetic resonance (NMR) characterization of compounds which were discussed in the manuscript with associated bioactivity data.

All COVID Moonshot compounds are publicly available as a screening collection that can be ordered in bulk or as singleton through Enamine. The compound identifiers of the COVID Moonshot collection are in the supplementary data files, together with Enamine's internal QC data comprising LC-MS spectra for all compounds and NMR spectra for selected compounds.

REFERENCES AND NOTES

1. N. Phillips, The coronavirus is here to stay – Here's what that means. *Nature* **590**, 382–384 (2021). doi: 10.1038/d41586-021-00396-2; pmid: 33594289
2. N. Winchester, "Covid-19 vaccinations: Is the Global South falling behind?" (UK House of Lords Library, 2021); https://lordslibrary.parliament.uk/covid-19-vaccinations-is-the-global-south-falling-behind/.
3. T. Zhao *et al*., Warnings regarding the potential coronavirus disease 2019 (COVID-19) transmission risk: Vaccination is not enough. *Infect. Control Hosp. Epidemiol.* **43**, 679–680 (2022). doi: 10.1017/ice.2021.63; pmid: 33563345
4. J. Hammond *et al*., Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *N. Engl. J. Med.* **386**, 1397–1408 (2022). doi: 10.1056/NEJMoa2118542; pmid: 35172054
5. M. W. McCarthy, Ensitrelvir as a potential treatment for COVID-19. *Expert Opin. Pharmacother.* **23**, 1995–1998 (2022). doi: 10.1080/14656566.2022.2146493; pmid: 36350029
6. A. Jayk Bernal *et al*., Molnupiravir for Oral Treatment of Covid-19 in Nonhospitalized Patients. *N. Engl. J. Med.* **386**, 509–520 (2022). doi: 10.1056/NEJMoa2116044; pmid: 34914868
7. E. de Wit, N. van Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: Recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016). doi: 10.1038/nrmicro.2016.81; pmid: 27344959
8. S. Ullrich, C. Nitsche, The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* **30**, 127377 (2020). doi: 10.1016/j.bmcl.2020.127377; pmid: 32738988
9. T. Pillaiyar, M. Manickam, V. Namasivayam, Y. Hayashi, S.-H. Jung, An Overview of Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *J. Med. Chem.* **59**, 6595–6628 (2016). doi: 10.1021/acs.jmedchem.5b01461; pmid: 26878082
10. L. Zhang *et al*., Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020). doi: 10.1126/science.abb3405; pmid: 32198291
11. Z. Jin *et al*., Structure of M^{pro} from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020). doi: 10.1038/s41586-020-2223-y; pmid: 32272481
12. D. R. Owen *et al*., An oral SARS-CoV-2 M^{pro} inhibitor clinical candidate for the treatment of COVID-19. *Science* **374**, 1586–1593 (2021). doi: 10.1126/science.abl4784; pmid: 34726479
13. H. Mukae *et al*., A Randomized Phase 2/3 Study of Ensitrelvir, a Novel Oral SARS-CoV-2 3C-Like Protease Inhibitor, in Japanese Patients with Mild-to-Moderate COVID-19 or Asymptomatic SARS-CoV-2 Infection: Results of the Phase 2a Part. *Antimicrob. Agents Chemother.* **66**, e0069722 (2022). doi: 10.1128/aac.00697-22; pmid: 36098519
14. H. Mukae *et al*., Efficacy and safety of ensitrelvir in patients with mild-to-moderate Coronavirus Disease 2019: The phase 2b part of a randomized, placebo-controlled, phase 2/3 study. *Clin. Infect. Dis.* **76**, 1403–1411 (2023). doi: 10.1093/cid/ciac933; pmid: 36477182
15. Study of PF-07321332 in Healthy Participants, ClinicalTrials.gov ID NCT04756531 (NIH, 2022); https://clinicaltrials.gov/study/NCT04756531.
16. Study of PBI-0451 in Healthy Subjects, ClinicalTrials.gov ID NCT05011812 (NIH, 2022); https://clinicaltrials.gov/study/NCT05011812.
17. J. Heskin *et al*., Caution required with use of ritonavir-boosted PF-07321332 in COVID-19 management. *Lancet* **399**, 21–22 (2022). doi: 10.1016/S0140-6736(21)02657-X; pmid: 34973713
18. F. von Delft *et al*., A white-knuckle ride of open COVID drug discovery. *Nature* **594**, 330–332 (2021). doi: 10.1038/d41586-021-01571-1; pmid: 34127864
19. J. Chodera, A. A. Lee, N. London, F. von Delft, Crowdsourcing drug discovery for pandemics. *Nat. Chem.* **12**, 581 (2020). doi: 10.1038/s41557-020-0496-2; pmid: 32555379
20. A. Douangamath *et al*., Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat. Commun.* **11**, 5047 (2020). doi: 10.1038/s41467-020-18709-w; pmid: 33028810
21. Main protease structure and XChem fragment screen (Diamond Light Source, 2020); https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html.
22. XChem @ Diamond; https://fragalysis.diamond.ac.uk/viewer/react/preview/target/Mpro.
23. D. C. Rees, M. Congreve, C. W. Murray, R. Carr, Fragment-based lead discovery. *Nat. Rev. Drug Discov.* **3**, 660–672 (2004). doi: 10.1038/nrd1467; pmid: 15286733
24. D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, H. Jhoti, Twenty years on: The impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* **15**, 605–619 (2016). doi: 10.1038/nrd.2016.109; pmid: 27417849
25. S. Cooper *et al*., Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010). doi: 10.1038/nature09304; pmid: 20686574
26. J. Lee *et al*., RNA design rules from a massive open laboratory. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2122–2127 (2014). doi: 10.1073/pnas.1313039111; pmid: 24469816
27. O. B. Cox *et al*., A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain. *Chem. Sci.* **7**, 2322–2330 (2016). doi: 10.1039/C5SC03115J; pmid: 29910922
28. A. Carbery, R. Skyner, F. von Delft, C. M. Deane, Fragment Libraries Designed to Be Functionally Diverse Recover Protein Binding Information More Efficiently Than Standard Structurally Diverse Libraries. *J. Med. Chem.* **65**, 11404–11413 (2022). doi: 10.1021/acs.jmedchem.2c01004; pmid: 35960886
29. N. Kitamura *et al*., Expedited Approach toward the Rational Design of Noncovalent SARS-CoV-2 Main Protease Inhibitors. *J. Med. Chem.* **65**, 2848–2865 (2022). doi: 10.1021/acs.jmedchem.1c00509; pmid: 33891389

30. D. Zaidman *et al*., An automatic pipeline for the design of irreversible derivatives identifies a potent SARS-CoV-2 M^{pro} inhibitor. *Cell Chem. Biol.* **28**, 1795–1806.e5 (2021). doi: 10.1016/j.chembiol.2021.05.018; pmid: 34174194
31. S. H. Han *et al*., Structure-Based Optimization of ML300-Derived, Noncovalent Inhibitors Targeting the Severe Acute Respiratory Syndrome Coronavirus 3CL Protease (SARS-CoV-2 $3CL^{pro}$). *J. Med. Chem.* **65**, 2880–2904 (2022). doi: 10.1021/acs.jmedchem.1c00598; pmid: 34347470
32. P. Schwaller *et al*., Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019). doi: 10.1021/acscentsci.9b00576; pmid: 31572784
33. A. Morris *et al*., Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed *de novo* design model. *Chem. Commun.* **57**, 5909–5912 (2021). doi: 10.1039/D1CC00050K; pmid: 34008627
34. A. S. J. S. Mey *et al*., Best Practices for Alchemical Free Energy Calculations. arXiv:2008.03067 [q-bio.BM] (2020).
35. R. Abel, L. Wang, D. L. Mobley, R. A. Friesner, A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr. Top. Med. Chem.* **17**, 2577–2585 (2017). doi: 10.2174/1568026617666170414142131; pmid: 28413950
36. K. A. Armacost, S. Riniker, Z. Cournia, Exploring Novel Directions in Free Energy Calculations. *J. Chem. Inf. Model.* **60**, 5283–5286 (2020). doi: 10.1021/acs.jcim.0c01266; pmid: 33222441
37. M. Shirts, V. S. Pande, Screen Savers of the World Unite! *Science* **290**, 1903–1904 (2000). doi: 10.1126/science.290.5498.1903; pmid: 17742054
38. M. I. Zimmerman *et al*., SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.* **13**, 651–659 (2021). doi: 10.1038/s41557-021-00707-0; pmid: 34031561
39. Y. Qiu *et al*., Development and Benchmarking of Open Force Field v1.0.0, the Parsley Small Molecule Force Field, ChemRxiv [Preprint] (2020); https://doi.org/10.26434/chemrxiv-2021-l070l-v4.
40. D. A. Rufa *et al*., Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning / molecular mechanics potentials. bioRxiv 2020.07.29.227959 [Preprint] (2020); https://doi.org/10.1101/2020.07.29.227959.
41. D. A. Rufa *et al*., Perses, version 0.10.2, Zenodo (2023); https://doi.org/10.5281/zenodo.8092391.
42. I. Zhang *et al*., Identifying and Overcoming the Sampling Challenges in Relative Binding Free Energy Calculations of a Model Protein:Protein Complex. *J. Chem. Theory Comput.* **19**, 4863–4882 (2023). doi: 10.1021/acs.jctc.3c00333; pmid: 37450482
43. P. Eastman *et al*., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017). doi: 10.1371/journal.pcbi.1005659; pmid: 28746339
44. A. Buitrago Santanilla *et al*., Organic chemistry. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015). doi: 10.1126/science.1259203; pmid: 25554781
45. D. Perera *et al*., A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018). doi: 10.1126/science.aap9112; pmid: 29371464
46. N. J. Gesmundo *et al*., Nanoscale synthesis and affinity ranking. *Nature* **557**, 228–232 (2018). doi: 10.1038/s41586-018-0056-8; pmid: 29686415
47. R. P. Thomas *et al*., A direct-to-biology high-throughput chemistry approach to reactive fragment screening. *Chem. Sci.* **12**, 12098–12106 (2021). doi: 10.1039/D1SC03551G; pmid: 34667575
48. P. Gehrtz *et al*., Optimization of Covalent MKK7 Inhibitors *via* Crude Nanomole-Scale Libraries. *J. Med. Chem.* **65**, 10341–10356 (2022). doi: 10.1021/acs.jmedchem.1c02206; pmid: 35912476
49. J.-Q. Chen, J.-H. Li, Z.-B. Dong, A Review on the Latest Progress of Chan-Lam Coupling Reaction. *Adv. Synth. Catal.* **362**, 3311–3331 (2020). doi: 10.1002/adsc.202000495
50. E. Resnick *et al*., Rapid Covalent-Probe Discovery by Electrophile-Fragment Screening. *J. Am. Chem. Soc.* **141**, 8951–8968 (2019). doi: 10.1021/jacs.9b02822; pmid: 31060360
51. M. N. L. Nalam *et al*., Evaluating the substrate-envelope hypothesis: Structural analysis of novel HIV-1 protease inhibitors designed to be robust against drug resistance. *J. Virol.* **84**, 5368–5378 (2010). doi: 10.1128/JVI.02531-09; pmid: 20237088
52. J. Jansen *et al*., SARS-CoV-2 infects the human kidney and drives fibrosis in kidney organoids. *Cell Stem Cell* **29**, 217–231.e8 (2021). doi: 10.1016/j.stem.2021.12.010; pmid: 35032430
53. Y. Unoh *et al*., Discovery of S-217622, a Noncovalent Oral SARS-CoV-2 3CL Protease Inhibitor Clinical Candidate for Treating COVID-19. *J. Med. Chem.* **65**, 6499–6512 (2022). doi: 10.1021/acs.jmedchem.2c00117; pmid: 35352927
54. "COVID Moonshot funded by COVID-19 Therapeutics Accelerator to rapidly develop a safe, globally accessible and affordable antiviral pill," Drugs for Neglected Diseases Initiative (DNDi) (2021); https://dndi.org/press-releases/2021/covid-moonshot-funded-by-wellcome-to-rapidly-develop-safe-globally-accessible-affordable-antiviral-pill/.
55. J. Strovel *et al*., "Early Drug Discovery and Development Guidelines: For Academic Researchers, Collaborators, and Start-up Companies," in *Assay Guidance Manual*, S. Markossian *et al*., Eds. (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2016).
56. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). doi: 10.1038/35057062; pmid: 11237011
57. A. R. Williamson, Creating a structural genomics consortium. *Nat. Struct. Biol.* **7**, 953 (2000). doi: 10.1038/80726; pmid: 11103997
58. J. Kaiser, "U.S. Cancer Institute 'Megaproject' to Target Common Cancer-Driving Protein," *ScienceInsider*, 24 June 2013; https://www.science.org/content/article/us-cancer-institute-megaproject-target-common-cancer-driving-protein.
59. A. Lee, J. Chodera, F. von Delft, "Why we are developing a patent-free Covid antiviral therapy," *Knowable Magazine*, 27 September 2021; https://doi.org/10.1146/knowable-092721-1.
60. S. Legare, F. Heide, B. A. Bailey-Elkin, J. Stetefeld, Improved SARS-CoV-2 main protease high-throughput screening assay using a 5-carboxyfluorescein substrate. *J. Biol. Chem.* **298**, 101739 (2022). doi: 10.1016/j.jbc.2022.101739; pmid: 35182525
61. M. G. Acker, D. S. Auld, Considerations for the design and reporting of enzyme assays in high-throughput screening applications. *Perspect. Sci.* **1**, 56–73 (2014). doi: 10.1016/j.pisc.2013.12.001
62. T. R. Malla *et al*., Mass spectrometry reveals potential of β-lactams as SARS-CoV-2 M^{pro} inhibitors. *Chem. Commun.* **57**, 1430–1433 (2021). doi: 10.1039/D0CC06870E; pmid: 33462575
63. G. D. Noske *et al*., A Crystallographic Snapshot of SARS-CoV-2 Main Protease Maturation Process. *J. Mol. Biol.* **433**, 167118 (2021). doi: 10.1016/j.jmb.2021.167118; pmid: 34174328
64. G. Winter *et al*., *DIALS*: Implementation and evaluation of a new integration package. *Acta Cryst.* **D74**, 85–97 (2018). doi: 10.1107/S2059798317017235; pmid: 29533234
65. G. Winter *et al*., How best to use photons. *Acta Cryst.* **D75**, 242–261 (2019). doi: 10.1107/S2059798319003528; pmid: 30950396
66. G. Winter, K. E. McAuley, Automated data collection for macromolecular crystallography. *Methods* **55**, 81–93 (2011). doi: 10.1016/j.ymeth.2011.06.010; pmid: 21763424
67. W. Kabsch, Integration, scaling, space-group assignment and post-refinement. *Acta Cryst.* **D66**, 133–144 (2010). doi: 10.1107/S0907444909047374; pmid: 20124693
68. G. Winter, C. M. C. Lobley, S. M. Prince, Decision making in *xia2*. *Acta Cryst.* **D69**, 1260–1273 (2013). doi: 10.1107/S0907444913015308; pmid: 23793152
69. C. Vonrhein *et al*., Data processing and analysis with the *autoPROC* toolbox. *Acta Cryst.* **D67**, 293–302 (2011). doi: 10.1107/S0907444911007773; pmid: 21460447
70. T. Krojer *et al*., The *XChemExplorer* graphical workflow tool for routine or large-scale protein–ligand structure determination. *Acta Cryst.* **D73**, 267–278 (2017). doi: 10.1107/S2059798316020234; pmid: 28291762
71. N. M. Pearce *et al*., A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nat. Commun.* **8**, 15123 (2017). doi: 10.1038/ncomms15123; pmid: 28436492
72. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of *Coot*. *Acta Cryst.* **D66**, 486–501 (2010). doi: 10.1107/S0907444910007493; pmid: 20383002
73. F. Long *et al*., *AceDRG*: A stereochemical description generator for ligands. *Acta Cryst.* **D73**, 112–122 (2017). doi: 10.1107/S2059798317000067; pmid: 28177307
74. T. Gorrie-Stone, xchem/XChemReview: Jan_2021, Zenodo (2022); https://doi.org/10.5281/zenodo.5939113.
75. I. J. Bruno *et al*., Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **44**, 2133–2144 (2004). doi: 10.1021/ci049780b; pmid: 15554684
76. S. J. Cottrell, T. S. G. Olsson, R. Taylor, J. C. Cole, J. W. Liebeschuetz, Validating and understanding ring conformations using small molecule crystallographic data. *J. Chem. Inf. Model.* **52**, 956–962 (2012). doi: 10.1021/ci200439d; pmid: 22372622
77. V. M. Corman *et al*., Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* **25**, 2000045 (2020). doi: 10.2807/1560-7917.ES.2020.25.3.2000045; pmid: 31992387
78. F. S. Varghese *et al*., Berberine and Obatoclax Inhibit SARS-Cov-2 Replication in Primary Human Nasal Epithelial Cells In Vitro. *Viruses* **13**, 282 (2021). doi: 10.3390/v13020282; pmid: 33670363
79. D. Jochmans, P. Leyssen, J. Neyts, A novel method for high-throughput screening to quantify antiviral activity against viruses that induce limited CPE. *J. Virol. Methods* **183**, 176–179 (2012). doi: 10.1016/j.jviromet.2012.04.011; pmid: 22575574
80. R. Rosales *et al*., Nirmatrelvir, Molnupiravir, and Remdesivir maintain potent in vitro activity against the SARS-CoV-2 Omicron variant. bioRxiv 2022.01.17.476685 [Preprint] (2022); https://doi.org/10.1101/2022.01.17.476685.
81. V. Gapsys *et al*., Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **11**, 1140–1152 (2020). doi: 10.1039/C9SC03754C; pmid: 34084371
82. J. A. Maier *et al*., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015). doi: 10.1021/acs.jctc.5b00255; pmid: 26574453
83. I. S. Joung, T. E. Cheatham3rd, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008). doi: 10.1021/jp8001614; pmid: 18593145
84. I. S. Joung, T. E. Cheatham 3rd, Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B* **113**, 13279–13290 (2009). doi: 10.1021/jp902584c; pmid: 19757835
85. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983). doi: 10.1063/1.445869
86. H. Xu, Optimal Measurement Network of Pairwise Differences. *J. Chem. Inf. Model.* **59**, 4720–4728 (2019). doi: 10.1103/PhysRevE.61.2361
87. G. E. Crooks, Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E* **61**, 2361–2366 (2000). doi: 10.1103/PhysRevE.61.2361
88. C. H. Bennett, Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22**, 245–268 (1976). doi: 10.1016/0021-9991(76)90078-4
89. S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, M. Schroeder, PLIP: Fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **43**, W443–W447 (2015). doi: 10.1093/nar/gkv315; pmid: 25873628
90. D. W. Kneller *et al*., Structural plasticity of SARS-CoV-2 3CL M^{pro} active site cavity revealed by room temperature X-ray crystallography. *Nat. Commun.* **11**, 3202 (2020). doi: 10.1038/s41467-020-16954-7; pmid: 32581217
91. M. Robinson, M. Henry, A. Morris, asapdiscovery/COVID_moonshot_submissions: Initial release for zenodo, version v0.1, Zenodo (2023); https://doi.org/10.5281/zenodo.8196348.
92. J. Scheen, asapdiscovery/COVID_moonshot_FECs_data: First Release, version 1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.8215553.
93. D. Fearon *et al*., Crystal structures of SARS-CoV-2 main protease screened against COVID Moonshot compounds by X-ray Crystallography at the XChem facility of Diamond Light Source, data set, Zenodo (2023); https://doi.org/10.5281/zenodo.8214528.

ACKNOWLEDGMENTS

The COVID Moonshot acknowledges funding by the Wellcome Trust on behalf of the Covid-19 Therapeutics Accelerator. The COVID Moonshot project is particularly grateful to UCB Pharma Ltd. and UCB SA for the support from their Medicinal and Computational Chemistry groups, to the Novartis Institute for Biomedical Research for generous in-kind ADME and PK contributions, to Takeda for in-kind contribution of antiviral assays/pan-corona biochemical assays, and to Nanosyn for protease panel assays. We thank CDD Vault and OpenEye Scientific for their in-kind contribution allowing the consortium to use their software. We also thank the numerous volunteers that contributed compound designs to the COVID Moonshot, the citizen scientist volunteers of Folding@home for donating their computing resources, and Amazon Web Services for key support of Folding@home infrastructure. **Funding:** Funding acknowledgments for each author are listed in the consortium spreadsheet (data S7).

Author contributions: M. L. Boby, D. Fearon, M. Ferla, M. Filep, L. Koekemoer, and M. C. Robinson, who contributed equally to this work, are listed alphabetically in the byline. M. Boby: Investigation, Writing – review & editing, Formal analysis, Software, and Visualization; J. Chodera: Writing – original draft, Conceptualization, Investigation, Writing – review & editing, Methodology, Resources, Funding acquisition, Data curation, Validation, Supervision, Formal analysis, Software, Project administration, and Visualization; A. von Delft: Writing – original draft, Conceptualization, Investigation, Writing – review & editing, Funding acquisition, Data curation, Validation, Supervision, Project administration, and Visualization; F. von Delft: Conceptualization, Writing – review & editing, Funding acquisition, Validation, Supervision, Project administration, and Visualization; D. Fearon: Writing – original draft, Conceptualization, Investigation, Writing – review & editing, Methodology, Resources, Data curation, Validation, Supervision, Formal analysis, Project administration, and Visualization; M. Ferla: Conceptualization, Investigation, Methodology, Data curation, Formal analysis, Software, and Visualization; M. Filep: Writing – original draft, Conceptualization, Investigation, Methodology, Resources, Formal analysis, and Visualization; L. Koekemoer: Writing – original draft, Investigation, Data curation, Validation, Formal analysis, and Project administration; A. A. Lee: Writing – original draft, Conceptualization, Investigation, Writing – review & editing, Methodology, Funding acquisition, Validation, Supervision, Formal analysis, Software, Project administration, and Visualization; N. London: Writing – original draft, Conceptualization, Investigation, Writing – review & editing, Methodology, Resources, Funding acquisition, Data curation, Validation, Supervision, Project administration, and Visualization; M. C. Robinson: Writing – original draft, Conceptualization, Investigation, Writing – review & editing, Methodology, Resources, Data curation, Validation, Supervision, Formal analysis, Software, Project administration, and Visualization. All other authors are listed as part of the COVID Moonshot Consortium, and the corresponding authors have confirmed that they meet the criteria for authorship, including data collection and analysis, development of methodology, and supplying resources. **Competing interests:** Disclosures for each author are listed individually in the consortium spreadsheet (data S7). The authors declare no additional competing interests. **Data and materials availability:** All compound designs, datasets, and x-ray structures can be browsed on the COVID Moonshot website (https://postera.ai/moonshot/). The compound submissions and experimental data are available via GitHub (https://github.com/postera-ai/COVID_moonshot_submissions), and the bioactivity data can be interactively browsed (https://covid.postera.ai/covid/activity_data). All data are also available in a permanent archive on Zenodo (*91*). Alchemical free-energy calculations code and datasets are indexed on GitHub (https://github.com/foldingathome/covid-moonshot) and are stored in a permanent archive on Zenodo (*92*). Analysis of alchemical free-energy calculations versus experimental potencies is stored at https://github.com/asapdiscovery/COVID_moonshot_FECs_data. All x-ray structures are available for interactive viewing and comparison or bulk download via Fragalysis (https://fragalysis.diamond.ac.uk/viewer/react/preview/target/Mpro). Structures were deposited to the PDB (data S4) and are also available in a permanent archive on Zenodo (*93*). With regard to synthesized compounds, we have made all compounds assayed here available from the current Enamine catalog and readily available for purchase from Enamine (and other suppliers) via the Manifold platform, accessible for each compound page on the COVID Moonshot website (https://covid.postera.ai/covid/activity_data).
License information: Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse.
This research was funded in whole or in part by the Wellcome Trust (grant no. 209407/Z/17/Z), a cOAlition S organization. The author will make the Author Accepted Manuscript (AAM) version available under a CC BY public copyright license.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo7201
Supplementary Text
Figs. S1 to S18
Table S1
References (*94*–*95*)
MDAR Reproducibility Checklist
Data S1 to S7

The COVID Moonshot Consortium

Hagit Achdout[1], Anthony Aimon[2,3], Dominic S. Alonzi[4], Robert Arbon[5], Jasmin C. Aschenbrenner[2,3], Blake H. Balcomb[2,3], Elad Bar-David[1], Haim Barr[6], Amir Ben-Shmuel[1], James Bennett[7,8], Vitaliy A. Bilenko[9,10], Melissa L. Boby[5,11], Bruce Borden[12], Pascale Boulet[13], Gregory R. Bowman[14], Lennart Brewitz[15], Juliane Brun[4], Sarma BVNBS[16], Mark Calmiano[17], Anna Carbery[2,18], Daniel W. Carney[19], Emma Cattermole[4], Edcon Chang[19], Eugene Chernyshenko[9], John D. Chodera[5], Austin Clyde[20], Joseph E. Coffland[21], Galit Cohen[6], Jason C. Cole[22], Alessandro Contini[23], Lisa Cox[24], Tristan Ian Croll[25,26], Milan Cvitkovic[27], Steven De Jonghe[28], Alex Dias[2,3], Kim Donckers[28], David L. Dotson[29], Alice Douangamath[2,3], Shirly Duberstein[6], Tim Dudgeon[30], Louise E. Dunnett[2,3], Peter Eastman[31], Noam Erez[1], Charles J. Eyermann[32], Michael Fairhead[7], Gwen Fate[33], Daren Fearon[2,3], Oleg Fedorov[7,8], Matteo Ferla[18], Rafaela S. Fernandes[34], Lori Ferrins[32], Mihajlo Filep[35], Richard Foster[36], Holly Foster[36,37], Laurent Fraisse[13], Ronen Gabizon[35], Adolfo García-Sastre[38,39,40,41,42], Victor O. Gawriljuk[34,43], Paul Gehrtz[35,44], Carina Gileadi[7], Charline Giroud[7,8], William G. Glass[5,37], Robert C. Glen[45], Itai Glinert[1], Andre S. Godoy[34], Marian Gorichko[10], Tyler Gorrie-Stone[2,3], Ed J. Griffen[46], Amna Haneef[47], Storm Hassell Hart[48], Jag Heer[49], Michael Henry[5], Michelle Hill[4,50], Sam Horrell[2,3], Qiu Yu Judy Huang[51], Victor D. Huliak[9], Matthew F. D. Hurley[52], Tomer Israely[1], Andrew Jajack[27], Jitske Jansen[53], Eric Jnoff[54], Dirk Jochmans[28], Tobias John[15,55], Benjamin Kaminow[5,56], Lulu Kang[57], Anastassia L. Kantsadi[4,58], Peter W. Kenny[59], J. L. Kiappes[4, 60], Serhii O. Kinakh[9], Lizbé Koekemoer[7], Boris Kovar[61], Tobias Krojer[7,62], Van Ngoc Thuy La[47], Sophie Laghnimi-Hahn[13], Alpha Lee[27], Bruce A. Lefker[33], Haim Levy[1], Ryan M. Lithgo[2,3], Ivan G. Logvinenko[9], Nir London[35], Petra Lukacik[2,3], Hannah Bruce Macdonald[5,63], Elizabeth M. MacLean[7], Laetitia L. Makower[4], Tika R. Malla[7], Peter G. Marples[2,3], Tatiana Matviiuk[9], Willam McCorkindale[64,63], Briana L. McGovern[38,39], Sharon Melamed[1], Kostiantyn P. Melnykov[9,10], Oleg Michurin[9], Pascal Miesen[65], Halina Mikolajek[2,3], Bruce F. Milne[66,67], David Minh[68], Aaron Morris[27], Garrett M. Morris[18], Melody Jane Morwitzer[69], Demetri Moustakas[70], Charles E. Mowbray[13], Aline M. Nakamura[34,71], Jose Brandao Neto[2,3], Johan Neyts[28], Luong Nguyen[27], Gabriela D. Noske[34], Vladas Oleinikovas[17,72], Glaucius Oliva[34], Gijs J. Overheul[65], C. David Owen[2,3], Ruby Pai[27], Jin Pan[27], Nir Paran[1], Alexander Matthew Payne[5,56], Benjamin Perry[13,73], Maneesh Pingle[16], Jakir Pinjari[16,74], Boaz Politi[1], Ailsa Powell[2,3], Vladimír Pšenák[61], Iván Pulido[5], Reut Puni[1], Victor L. Rangel[75,76], Rambabu N. Reddi[35], Paul Rees[77], St Patrick Reid[69], Lauren Reid[46], Efrat Resnick[35], Emily Grace Ripka[27], Matthew C. Robinson[27], Ralph P. Robinson[33], Jaime Rodriguez-Guerra[78], Romel Rosales[38,39], Dominic A. Rufa[5,56], Kadi Saar[64], Kumar Singh Saikatendu[19], Eidarus Salah[15], David Schaller[78], Jenke Scheen[5], Celia A. Schiffer[51], Christopher J. Schofield[15], Mikhail Shafeev[9], Aarif Shaikh[16], Ala M. Shaqra[51], Jiye Shi[54,79], Khriesto Shurrush[6], Sukrit Singh[5], Assa Sittner[1], Peter Sjö[13], Rachael Skyner[2,3], Adam Smalley[17], Bart Smeets[80], Mihaela D. Smilova[7], Leonardo J. Solmesky[6], John Spencer[48], Claire Strain-Damerell[2,3], Vishwanath Swamy[16,81], Hadas Tamir[1], Jenny C. Taylor[82], Rachael E. Tennant[83], Warren Thompson[2,3], Andrew Thompson[7,84], Susana Tomásio[85], Charles W. E. Tomlinson[2,3], Igor S. Tsurupa[9], Anthony Tumber[15], Ioannis Vakonakis[4,86], Ronald P. van Rij[65], Laura Vangeel[28], Finny S. Varghese[65,87], Mariana Vaschetto[85], Einat B. Vitner[1], Vincent Voelz[52], Andrea Volkamer[78,88], Frank von Delft[2,3,89], Annette von Delft[7], Martin A. Walsh[2,3], Walter Ward[90], Charlie Weatherall[91], Shay Weiss[1], Kris M. White[38,39], Conor Francis Wild[2,3], Karolina D. Witt[92], Matthew Wittmann[5], Nathan Wright[7], Yfat Yahalom-Ronen[1], Nese Kurt Yilmaz[51], Daniel Zaidmann[35], Ivy Zhang[5], Hadeer Zidane[6], Nicole Zitzmann[4], Sarah N. Zvornicanin[51]

[1]Israel Institute for Biological Research, Department of Infectious Diseases, Ness-Ziona, Israel. [2]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot, OX11 0DE, UK. [3]Research Complex at Harwell, Harwell Science and Innovation Campus, Didcot OX11 0FA, UK. [4]University of Oxford, Department of Biochemistry, Oxford Glycobiology Institute, South Parks Road, Oxford OX1 3QU, UK. [5]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, Computational and Systems Biology Program, New York, NY 10065, USA. [6]The Weizmann Institute of Science, Wohl Institute for Drug Discovery of the Nancy and Stephen Grand Israel National Center for Personalized Medicine, Rehovot, 7610001, Israel. [7]University of Oxford, Nuffield Department of Medicine, Centre for Medicines Discovery, Oxford, OX3 7DQ, UK. [8]University of Oxford, Nuffield Department of Medicine, Target Discovery Institute, Oxford, OX3 7FZ, UK. [9]Enamine Ltd, Kyiv, 02094, Ukraine. [10]Taras Shevchenko National University of Kyiv, Kyiv, 01601, Ukraine. [11]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, Pharmacology Graduate Program, New York, NY 10065, USA. [12]Folding@home Consortium. [13]Drugs for Neglected Diseases Initiative (DNDi), Geneva, 1202, Switzerland. [14]University of Pennsylvania, Departments of Biochemistry and Biophysics and Bioengineering, Philadelphia, PA 19083, USA. [15]University of Oxford, Department of Chemistry, Chemistry Research Laboratory, Oxford, OX1 3TA, UK. [16]Sai Life Sciences Limited, ICICI Knowledge Park, Shameerpet, Hyderabad 500 078, Telangana, India. [17]UCB, Slough, SL1 3WE, UK. [18]University of Oxford, Department of Statistics, Oxford OX1 3LB, UK. [19]Takeda Development Center Americas, Inc., San Diego, CA 92121, USA. [20]Argonne National Lab, Lemont, IL 60439, USA. [21]Cauldron Development Oy, Helsinki, 00140, Finland. [22]Cambridge Crystallographic Data Centre, Cambridge, CB2 1EZ, UK. [23]University of Milan, Department of General and Organic Chemistry, Milan, 20133, Italy. [24]Life Compass Consulting Ltd, Macclesfield, SK10 5UE, UK. [25]The University of Cambridge, Cambridge Institute for Medical Research, Department of Haematology, Cambridge CB2 0XY, UK. [26]Present address: Altos Labs, BioML group, Great Abington, CB21 6GP. [27]PostEra Inc., Cambridge, MA, 02142, USA. [28]KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Virology and Chemotherapy, Leuven, Belgium. [29]Datryllic LLC, Phoenix AZ, 85003, USA. [30]Informatics Matters Ltd, Bicester, OX26 6JU, UK. [31]Stanford University, Department of Chemistry, Stanford, CA 94305, USA. [32]Northeastern University, Department of Chemistry and Chemical Biology, Boston MA 02115, USA. [33]Thames Pharma Partners LLC, Mystic, CT 06355, USA. [34]University of Sao Paulo, Sao Carlos Institute of Physics, Sao Carlos, 13563-120, Brazil. [35]The Weizmann Institute of Science, Department of Chemical and Structural Biology, Rehovot, 7610001, Israel. [36]University of Leeds, School of Chemistry, Leeds, LS2 9JT, UK. [37]Present address: Exscientia, Oxford Science Park, Oxford, OX4 4GE, UK. [38]Icahn School of Medicine at Mount Sinai, Department of Microbiology, New York, NY 10029, USA. [39]Icahn School of Medicine at Mount Sinai, Global Health and Emerging Pathogens Institute, New York, NY 10029, USA. [40]Icahn School of Medicine at Mount Sinai, Department of Medicine, Division of Infectious Diseases, New York, NY 10029, USA. [41]Icahn School of Medicine at Mount Sinai, The Tisch Cancer Institute, New York, NY 10029, USA. [42]Icahn School of Medicine at Mount Sinai, Department of Pathology, Molecular and Cell-Based Medicine, New York, NY 10029, USA. [43]Present address: University of Groningen, Groningen Research Institute of Pharmacy, Department of Drug Design, Groningen, 9700 AV, Netherlands. [44]Present address: Merck Healthcare KGaA, Darmstadt, 64293, Germany. [45]University of Cambridge, Department of Chemistry, Cambridge, CB2 1EW, UK. [46]MedChemica Ltd, Macclesfield, Cheshire. SK11 6PU UK. [47]Illinois Institute of Technology, Department of Biology, Chicago IL 60616 USA. [48]University of Sussex, Department of Chemistry, School of Life Sciences, Brighton, East Sussex, BN1 9QJ, UK. [49]Syngene International Limited, Headington, Oxford, OX3 7BZ, UK. [50]Present address: Sir William Dunn School of Pathology, Oxford. OX1 3RE, UK. [51]University of Massachusetts, Chan Medical School, Department of Biochemistry and Molecular Biotechnology, Worcester MA 01655, USA. [52]Temple University, Department of Chemistry, Philadelphia, PA 19122, USA. [53]RWTH Aachen University, Institute of Experimental Medicine and Systems Biology, Aachen, 52074, Germany. [54]UCB, Chemin du Foriest, 1420 Braine-l'Alleud, Belgium. [55]Present address: AMSilk, Neuried, 82061, Germany. [56]Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, Tri-Institutional Program in Computational Biology and Medicine, New York, NY 10065, USA. [57]Illinois Institute of Technology, Department of Applied Mathematics, Chicago IL 60616 USA. [58]University of Thessaly, Department of Biochemistry and Biotechnology, Larissa, 415 00, Greece. [59]Berwick-on-Sea, North Coast Road, Blanchisseuse, Saint George, Trinidad and Tobago. [60]Present address: University College of London, Department of Chemistry, London WC1H OAJ, UK. [61]M2M solutions s.r.o. Žilina, 010 01, Slovakia. [62]MAX IV Laboratory, Fotongatan 2, 224 84 Lund, Sweden. [63]Present address: Charm Therapeutics, London, N1C 4AG, UK. [64]University of Cambridge, Cavendish Laboratory, Cambridge, CB3 0HE UK. [65]Radboud University Medical Center, Department of Medical Microbiology, Radboud Institute for Molecular Life Sciences, Nijmegen, 6525 GA, Netherlands. [66]University of Aberdeen, Department of Chemistry, Old Aberdeen, AB24 3UE Scotland, UK. [67]University of Coimbra, CFisUC, Department of Physics, Coimbra, 3004-516, Portugal. [68]Illinois Institute of Technology, Department of Chemistry, Chicago IL 60616 USA. [69]University of Nebraska Medical Centre, Dept of Pathology and Microbiology, Omaha, NE 68198-5900, USA. [70]Relay Therapeutics, Cambridge, MA 02139, USA. [71]Present address: Instituto Butantan, Sao Paulo, 05503-900, Brazil. [72]Present address: Monte Rosa Therapeutics, Basel, CH 4057, Switzerland. [73]Present address: Medicxi, Geneva, 1204, Switzerland. [74]Present address: Sun Pharma Advanced Research Company (SPARC), Baroda, India. [75]University of São Paulo, Ribeirão Preto School of Pharmaceutical Sciences, Ribeirão Preto - SP/CEP 14040-903, Brazil. [76]Present

address: Evotec (UK) Ltd, Milton Park, Abingdon, Oxfordshire, OX14 4RZ, UK. [77]Compass Bussiness Partners Ltd, Southcliffe, Bucks, SL9 0PD, UK. [78]Charité – Universitätsmedizin Berlin, In silico Toxicology and Structural Bioinformatics, Berlin, 10117, Germany. [79]Present address: Eli Lilly and Company, San Diego, CA 92121, USA. [80]Radboud University Medical Center, Department of pathology, Radboud Institute for Molecular Life Sciences, Nijmegen, 6525 GA, Netherlands. [81]Present address: TCG Life Sciences, Pune, India. [82]University of Oxford, Nuffield Department of Medicine, Wellcome Centre for Human Genetics, Oxford OX3 7BN, UK. [83]Lhasa Limited, Leeds, LS11 5PS, UK. [84]Present address: Walter and Eliza Hall Institute, Parkville 3052, Victoria, Australia. [85]Collaborative Drug Discovery, Cambridge, CB2 1GE, UK. [86]Present address: Lonza Biologics, Lonza Ltd, Lonzastrasse, CH-3930 Visp, Switzerland. [87]Present address: uniQure Biopharma, Amsterdam, 1105 BP, Netherlands. [88]Present address: Saarland University, Data Driven Drug Design, Campus - E2.1, 66123 Saarbrücken, Germany. [89]University of Johannesburg, Department of Biochemistry, Auckland Park, 2006, South Africa. [90]Walter Ward Consultancy and Training, Derbyshire, SK22 4AA, UK. [91]Collaborative Drug Discovery, Burlingame, CA 94010, USA. [92]University of Oxford, Nuffield Department of Medicine, Pandemic Sciences Institute, Oxford, Oxon, OX3 7DQ, UK.

Submitted 21 February 2022; accepted 9 October 2023
10.1126/science.abo7201

RESEARCH ARTICLE SUMMARY

METABOLISM

Aster-dependent nonvesicular transport facilitates dietary cholesterol uptake

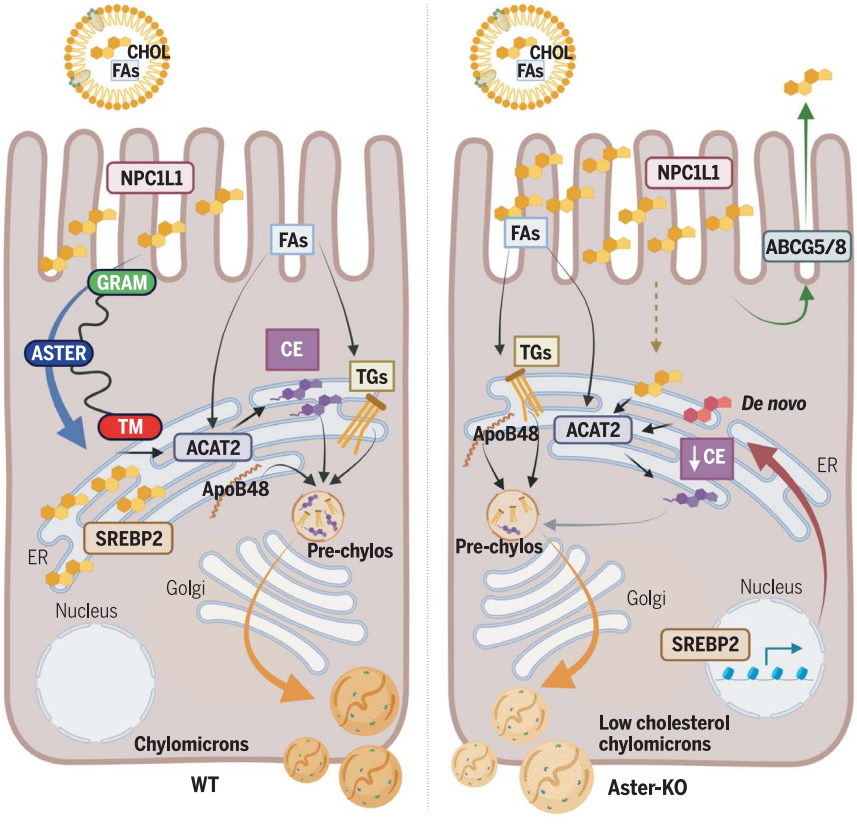
Alessandra Ferrari, Emily Whang, Xu Xiao, John P. Kennelly, Beatriz Romartinez-Alonso, Julia J. Mack, Thomas Weston, Kai Chen, Youngjae Kim, Marcus J. Tol, Lara Bideyan, Alexander Nguyen, Yajing Gao, Liujuan Cui, Alexander H. Bedard, Jaspreet Sandhu, Stephen D. Lee, Louise Fairall, Kevin J. Williams, Wenxin Song, Priscilla Munguia, Robert A. Russell, Martin G. Martin, Michael E. Jung, Haibo Jiang, John W. R. Schwabe, Stephen G. Young, Peter Tontonoz*

INTRODUCTION: Hypercholesterolemia is a major risk factor for cardiovascular disease. The small intestine is the gatekeeper of dietary cholesterol absorption and a tractable target for lipid-lowering therapies. Niemann-Pick C1 Like 1 (NPC1L1), the target of the hypocholesterolemic drug ezetimibe (EZ), facilitates the deposition of dietary cholesterol from the gut lumen into the apical plasma membrane (PM) of enterocytes. Subsequently, cholesterol moves to the endoplasmic reticulum (ER) where it is esterified by ACAT2. Cholesterol ester is packaged into chylomicrons for release into the circulation and delivery to tissues. How cholesterol moves from the brush border of enterocytes to the ER, however, has been a longstanding question.

RATIONALE: Prior studies have proposed that endosomal pathways, possibly involving NPC1L1, may contribute to the movement of cholesterol into enterocytes. However, the involvement of nonvesicular pathways has not been tested. Aster proteins are ER-resident nonvesicular transp… ers that bind cholesterol and facilitate its … moval from the PM by forming PM-ER contact sites. We addressed the role of Asters in intestinal physiology and systemic cholesterol homeostasis using cellular assays, intestinal enteroid cultures, structural biology, and genetically modified mouse models.

RESULTS: We discovered that Aster-B is translocated to the brush border of enterocytes of the small intestine in response to oral administration of cholesterol to mice. Since Aster-B and Aster-C are expressed in the small intestine, we generated mice with genetic deletions of both proteins. Loss of Aster-dependent nonvesicular cholesterol transport in enterocytes compromised dietary cholesterol absorption, despite the presence of NPC1L1. We used intestinal enteroid cultures to demonstrate that Aster deletion caused an accumulation of ALOD4-binding accessible cholesterol at the PM of enterocytes, indicative of impaired cholesterol movement to the ER. Mice lacking Asters also showed evidence of ER cholesterol depletion, including reduced cholesterol ester formation and activation of the SREBP2 transcriptional program for de novo cholesterol synthesis. Aster-deficient mice produced chylomicrons depleted of cholesterol esters and were protected from diet-induced hypercholesterolemia. Both Aster-B and Aster-C bound EZ, and we resolved the crystal structure of the Aster-C-EZ complex to 1.6 Å resolution. Based on this observation, we examined potential cooperation between Asters and NPC1L1. We showed that NPC1L1, by saturating the brush border with diet-derived cholesterol, was necessary for Aster recruitment to the apical PM of enterocytes. Acting downstream of NPC1L1, Asters were required to move cholesterol deposited by NPC1L1 to the enterocyte ER. Finally, the small-molecule Aster inhibitor AI-3d caused the accumulation of accessible cholesterol at the PM of both mouse and human intestinal enteroids and reduced dietary cholesterol absorption in mice.

CONCLUSION: Asters and NPC1L1 perform sequential, nonredundant functions in dietary cholesterol uptake. Nonvesicular cholesterol transport mechanisms downstream of NPC1L1 action at the PM are important for cholesterol movement in enterocytes and contribute to systemic sterol homeostasis. These findings highlight the Aster pathway as a physiologically important determinant of dietary lipid absorption that can be targeted pharmacologically. ■



Asters facilitate intracellular movement of dietary cholesterol in enterocytes. NPC1L1 mediates the deposition of dietary cholesterol at the brush border, thus recruiting Asters to form PM-ER contact sites. Asters subsequently facilitate the nonvesicular movement of cholesterol to ER for esterification. Deletion of Asters causes ER sterol depletion, activation of SREBP2, production of chylomicrons depleted of cholesterol esters, and reduced systemic cholesterol burden. [Figure created using BioRender]

The list of author affiliations is available in the full article.
*Corresponding author. Email: ptontonoz@mednet.ucla.edu
Cite this article as A. Ferrari *et al.*, *Science* **382**, eadf0966 (2023). DOI: 10.1126/science.adf0966

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adf0966

RESEARCH ARTICLE

METABOLISM

Aster-dependent nonvesicular transport facilitates dietary cholesterol uptake

Alessandra Ferrari[1,2]†, Emily Whang[1,2,3]†, Xu Xiao[1,2], John P. Kennelly[1,2], Beatriz Romartinez-Alonso[4], Julia J. Mack[5], Thomas Weston[5,6], Kai Chen[7,8], Youngjae Kim[9], Marcus J. Tol[1,2], Lara Bideyan[1,2], Alexander Nguyen[1,2,10], Yajing Gao[1,2], Liujuan Cui[1,2], Alexander H. Bedard[1,2], Jaspreet Sandhu[1,11], Stephen D. Lee[1,2], Louise Fairall[4], Kevin J. Williams[2,12], Wenxin Song[5,6], Priscilla Munguia[5,6], Robert A. Russell[13], Martin G. Martin[3], Michael E. Jung[9], Haibo Jiang[7,8], John W. R. Schwabe[4], Stephen G. Young[5,6], Peter Tontonoz[1,2]*

Intestinal absorption is an important contributor to systemic cholesterol homeostasis. Niemann-Pick C1 Like 1 (NPC1L1) assists in the initial step of dietary cholesterol uptake, but how cholesterol moves downstream of NPC1L1 is unknown. We show that Aster-B and Aster-C are critical for nonvesicular cholesterol movement in enterocytes. Loss of NPC1L1 diminishes accessible plasma membrane (PM) cholesterol and abolishes Aster recruitment to the intestinal brush border. Enterocytes lacking Asters accumulate PM cholesterol and show endoplasmic reticulum cholesterol depletion. Aster-deficient mice have impaired cholesterol absorption and are protected against diet-induced hypercholesterolemia. Finally, the Aster pathway can be targeted with a small-molecule inhibitor to manipulate cholesterol uptake. These findings identify the Aster pathway as a physiologically important and pharmacologically tractable node in dietary lipid absorption.

The intestine regulates systemic lipid homeostasis by gating dietary cholesterol intake (*1*). Cholesterol is absorbed by enterocytes and packaged into chylomicrons, which are released into the lymphatics and ultimately reach the systemic circulation (*2*). Most of the cholesterol in chylomicrons is esterified, and cholesterol ester (CE) is necessary for chylomicron packaging. Free cholesterol deposited into the inner leaflet of the apical plasma membrane (PM) by Niemann-Pick C1-Like 1 (NPC1L1) must subsequently move to the ER to be esterified by ACAT2 (*3*–*5*). NPC1L1 is an important mediator of intestinal cholesterol uptake (*6*, *7*), but how cholesterol is delivered to the enterocyte ER for esterification remains unknown.

NPC1L1 is a target of the drug ezetimibe (EZ) (*8*–*10*). The combination of EZ and a statin further reduces cardiovascular events compared with statin alone (*10*), validating the intestine as a therapeutic target for the regulation of cholesterol homeostasis. EZ was initially discovered to be an inhibitor of ACAT2 (*11*). Its notable effects on cholesterol absorption led to its FDA approval even before it was found that it inhibits NPC1L1. Genetic ablation of NPC1L1 or EZ administration impairs cholesterol absorption (*6*). Treatment with EZ protects mice from diet-induced hypercholesterolemia and atherosclerosis (*1*, *3*, *12*, *13*). Structural analyses have revealed that the N-terminal domain of NPC1L1 contains a cavity that accommodates cholesterol (*14*–*16*). Cholesterol moves through this channel to diffuse into the lipid bilayer. EZ binds within the channel, thereby blocking cholesterol deposition. Despite the importance of the NPC1L1-ACAT2 axis for cholesterol homeostasis, how cholesterol received by NPC1L1 at the PM ultimately reaches the ER is unknown (*17*–*24*).

Aster proteins (Aster-A, -B, and -C, encoded by *Gramd1a, Gramd1b,* and *Gramd1c*, respectively) bind cholesterol and facilitate its movement between membranes (*25*–*27*). Members of this family have a central cholesterol-binding pocket that is flanked by a GRAM domain at the N terminus and an ER transmembrane domain at the C terminus. The GRAM domain binds the PM in response to cholesterol loading, allowing Asters to transfer cholesterol down a concentration gradient from the PM to the ER. The role of nonvesicular cholesterol transport in cell function is highly cell-type specific. In mice, Aster-B is required for CE storage and corticosteroid synthesis in the adrenal cortex (*25*), whereas Aster-C is important for hepatic reverse cholesterol transport (*28*).

We show that NPC1L1 and Asters play sequential, nonredundant roles in the delivery of dietary cholesterol from the intestinal lumen to the enterocyte ER. Asters are recruited to the enterocyte PM upon cholesterol loading, and loss of their expression impairs CE production. Treatment of mice with a small-molecule Aster inhibitor reduces the systemic absorption of dietary cholesterol, and mice lacking Asters in the intestine are protected from diet-induced hypercholesterolemia. Our findings support a model in which the principal function of NPC1L1 is to enrich the enterocyte PM with dietary cholesterol. This enrichment facilitates the recruitment of Asters to the PM where they mediate nonvesicular cholesterol trafficking to the ER. These findings identify intestinal Asters as key players in dietary lipid absorption and potential targets for the control of cholesterol homeostasis.

Results

Intestinal Aster expression is regulated by liver x receptors

Transcripts for Aster-A, -B, and -C are expressed in the small intestine (SI) (Fig. 1A), and RNA-seq from proximal jejunal scrapings of C57BL6/J mice confirmed robust expression of *Gramd1b* and *Gramd1c* and low expression of *Gramd1a* (fig. S1A). The most abundant *Gramd1c* transcript in the SI encodes a truncated Aster-C lacking the N-terminal GRAM domain. RNA-seq also revealed the presence of two transcripts for *Gramd1b* in the intestine; one was the same as that identified previously in macrophages (*25*), while the other had an extended N-terminal region upstream of the GRAM domain (fig. S1A). An intestine-specific promoter in the *Gramd1b* locus (chr9: 40465470-40465756) defined the longer variant. Absolute quantification of *Gramd1* transcripts in intestinal samples confirmed that those for Aster-B and -C were most abundant (fig. S1B). Aster-B was expressed along the entire length of the SI (duodenum to the ileum) (fig. S1C). In Caco-2 cells, transcripts for Aster-B and -C were induced upon differentiation (fig. S1D). Liver x receptor transcription factors are key regulators of cholesterol metabolism (*29*, *30*). Treatment of wild-type (WT) mice with the LXR agonist GW3965 induced the expression of all three Asters in the SI (fig. S1E), consistent with a role for the Aster pathway in intestinal cholesterol flux.

[1]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [2]Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA. [3]Pediatric Gastroenterology, Hepatology, and Nutrition, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [4]Institute for Structural and Chemical Biology, University of Leicester, Leicester LE1 7RH, UK. [5]Department of Medicine, Division of Cardiology, University of California, Los Angeles, Los Angeles, CA, 90095, USA. [6]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA. [7]Department of Chemistry, The University of Hong Kong, Hong Kong 999077, China. [8]School of Molecular Sciences, The University of Western Australia, Crawley, WA 6009, Australia. [9]Department of Chemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA. [10]Vatche and Tamar Manoukian Division of Digestive Diseases, Department of Medicine David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [11]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA. [12]UCLA Lipidomics Core, University of California, Los Angeles, Los Angeles, CA 90095, USA. [13]National Deuteration Facility, Australian Nuclear Science and Technology Organisation, Lucas Heights, NSW 2234, Australia.
*Corresponding author. Email: ptontonoz@mednet.ucla.edu
†These authors contributed equally to this work.

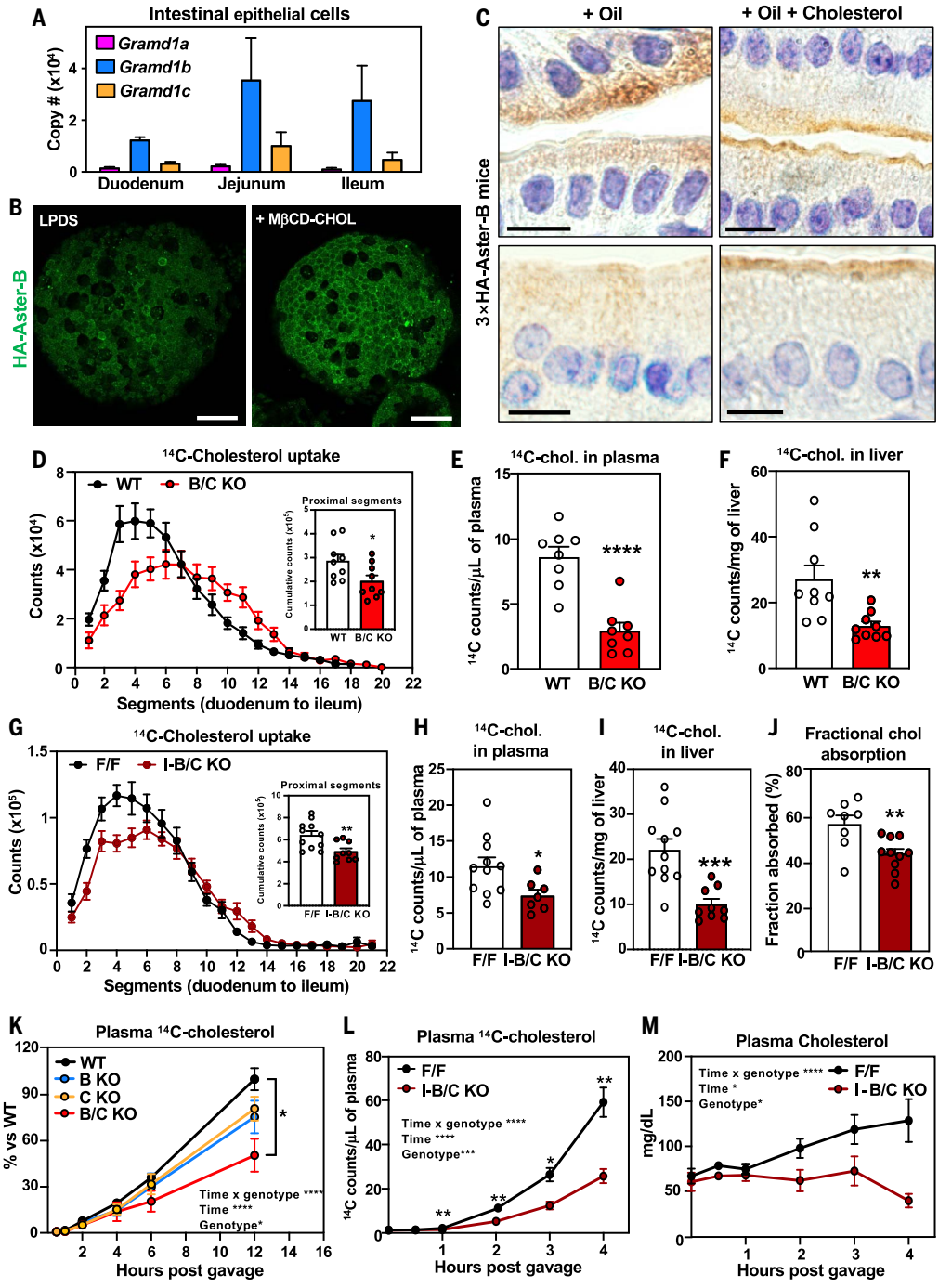


Fig. 1. Aster proteins modulate dietary cholesterol uptake. (**A**) Absolute quantification of *Gramd1a/b/c* mRNA in the duodenum, jejunum, and ileum of C57BL/6J male mice (n = 5). (**B**) IF microscopy of HA-Aster-B in intestinal organoids from 3×HA-Aster-B mice during sterol deprivation (left) or loading with MβCD-cholesterol (right). Scale bar is 50 µm. (**C**) Immunohistochemical staining of HA-Aster-B in small intestines from 3×HA-Aster-B mice after a gastric gavage with corn oil or corn oil with cholesterol. For upper and lower panels, scale bar is 20 and 10 µm, respectively. (**D**) Radioactivity in intestinal segments of female WT and B/C KO mice after oral gavage with olive oil containing [^{14}C] cholesterol for 2 hours (n = 9 per group), and cumulative counts in the proximal intestine. (**E**) Radioactivity in plasma of mice described in (D). (**F**) Radioactivity in livers of mice described in (D). (**G**) Radioactivity in intestinal segments of female F/F and I-B/C KO mice after an oral challenge of olive oil containing [^{14}C] cholesterol for 2 hours (n = 9 per group), and cumulative counts in the proximal intestine. (**H**) Radioactivity in plasma of mice described in (H). (**I**) Radioactivity in livers of mice described in (I). (**J**) Cholesterol absorption measured by the fecal dual-isotope ratio method (n = 8 to 10 per group). (**K**) Kinetics of radioactivity in plasma of female WT (n = 9), BKO (n = 5), CKO (n = 5), B/C KO (n = 5) mice after an oral challenge of olive oil containing [^{14}C]cholesterol. (**L**) Kinetics of radioactivity in plasma of female F/F and I-B/C KO mice after injection of Poloxamer-407 and an oral challenge of [^{14}C]cholesterol in olive oil. (**M**) Kinetics of total cholesterol in mice described in (L). Data are expressed as mean ± SEM. Statistical analysis: unpaired t test for (D, E, F, G, H, I, and J); 2-way analysis of variance (ANOVA) with Tukey's multiple comparisons test for (K); and 2-way ANOVA with Sidak's multiple comparisons test for (L and M). *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

Aster proteins facilitate cholesterol uptake in the small intestine

To visualize endogenous Aster-B movement in enterocytes, we used CRISPR-Cas9 editing to insert a 3×HA tag into the mouse *Gramd1b* locus (fig. S1, F and G). Immunohistochemistry of intestinal tissue revealed that HA-Aster-B was enriched in the villi of the jejunum with lower expression in the crypts (fig. S1H). To study Aster localization in culture we derived enteroids from intestinal crypts of HA-Aster-B mice. Confocal imaging demonstrated that Aster-B was recruited to the PM in response to loading with methyl-β-cyclodextrin (MβCD)–cholesterol (Fig. 1B). Accordingly, immunohistochemistry showed a greater distribution of HA-Aster-B at the brush border of enterocytes 1 hour after intra-gastric gavage with cholesterol in corn oil versus corn oil alone (Fig. 1C).

To explore the function of Asters in intestinal physiology, we generated global knockout mice for Aster-B (B-KO), Aster-C (C-KO), or both (B/C-KO) (fig. S2, A and B). Deletion of Asters was verified in jejunal scrapings (fig. S2, C, D, and E). The SI from B/C-KO mice had no obvious histological abnormalities (fig. S2F), and body weight and intestinal length showed minimal differences across genotypes (fig. S2G). To evaluate the role of Asters in cholesterol uptake, we administered [^{14}C]cholesterol by gastric gavage and measured radioactivity in SI 2 hours later. No differences in [^{14}C]cholesterol absorption were detected in single-KO mice (fig. S2, H to M), but we observed markedly reduced absorption in the proximal intestine of B/C-KO mice (Fig. 1D and fig. S2, N and O). B/C-KO mice also had reduced amounts of [^{14}C]cholesterol in the plasma and liver 2 hours after gavage (Fig. 1, E and F).

To confirm that this impairment in cholesterol absorption resulted from Aster deficiency in enterocytes, we generated mice with tamoxifen-inducible, intestine-specific deletion of Aster-B (I-B-KO), Aster-C (I-C-KO), or both (I-B/C-KO) by intercrossing double "floxed" mice with Villin-Cre^{ERT2} transgenic mice (fig. S3, A to F). I-B/C KO mice displayed normal crypt cell proliferation as revealed by OLFM4 staining (fig. S3G). Minimal differences in body weight or SI length were observed between control and I-B/C KO mice (fig. S3H). Thus, loss of Aster expression does not appear to compromise intestinal development.

Cholesterol uptake was similar in I-B KO, I-C-KO, and littermate control mice (fig. S3, I to N). However, I-B/C-KO mice had reduced [^{14}C]cholesterol absorption compared with floxed controls (similar to global B/C-KO mice) (Fig. 1, G to I). Cumulative counts in the proximal tract (Fig. 1G) were reduced, whereas counts in medial and distal regions were not different (fig. S3, O and P). Fractional cholesterol absorption, assessed by fecal dual isotope labeling, confirmed impaired cholesterol absorption in I-B/C-KO mice (Fig. 1J). By contrast, there were no detectable reductions in ^{14}C counts in intestinal segments, plasma, and liver of I-B/C-KO mice after oral gavage of [^{14}C]triolein (fig. S4, A to D). Glucose absorption was not affected in I-B/C-KO mice (fig. S4E). Thus, loss of Aster function in enterocytes selectively impairs cholesterol absorption.

Next, we followed the appearance of orally administered [^{14}C]cholesterol in the plasma over 12 hours. There was a marked reduction in ^{14}C counts in the plasma of B/C-KO mice, whereas B-KO and C-KO mice exhibited an intermediate phenotype (Fig. 1K). We also injected mice with the lipoprotein lipase inhibitor Poloxamer-407, gavaged them with [^{14}C]cholesterol, and collected plasma up to 4 hours post gavage. The appearance of [^{14}C]cholesterol in the plasma over time, as well as total plasma cholesterol levels, was reduced in I-B/C-KO mice (Fig. 1, L and M).

Nonvesicular cholesterol transport enables intestinal CE production

Adequate CE is required for chylomicron production by the intestine. We hypothesized that Aster-mediated nonvesicular transport from the PM to the ER facilitates cholesterol esterification and subsequent incorporation into chylomicrons. We used nanoscale secondary ion mass spectrometry (NanoSIMS) imaging to visualize cholesterol uptake and intracellular distribution by enterocytes (*31*, *32*). Mice received an intra-gastric gavage of [^{13}C]fatty acids and [^{2}H]cholesterol in olive oil. Biopsies from the duodenum were harvested 2 hours later. Backscattered electron images of duodenal sections verified the integrity of intestinal villi (Fig. 2A and fig. S5A). NanoSIMS images of the same sections (Fig. 2A and fig. S5A) revealed reduced amounts of ^{2}H in medial and distal segments of the duodenum in B/C-KO mice. Quantification of $^{13}C^-$ and $^{2}H^-$ secondary ions confirmed reduced cholesterol uptake by enterocytes lacking Asters (Fig. 2B and fig. S5B).

Previous studies showed that in macrophages (*33*) and hepatocytes (*28*) Asters specifically recognize the accessible pool of PM cholesterol (*34*, *35*). To prove that Asters transfer accessible cholesterol from the PM to ER in intestinal epithelial cells, we stained WT and Aster-B/C KO enteroids with ALOD4, a bacterial peptide that selectively binds to accessible cholesterol. Confocal microscopy revealed enhanced ALOD4 staining at the PM of Aster-B/C KO enteroids loaded with MβCD-cholesterol (Fig. 2C), consistent with reduced cholesterol movement to the ER. Accordingly, lipidomic analysis of jejunal scrapings revealed reduced CE in B/C-KO mice after feeding (Fig. 3A). Other lipid species were not affected by Aster deficiency (fig. S6A). Consistent with these findings, ^{14}C-labeled CE accumulation was reduced in the proximal jejunum of Aster-deficient mice 2 hours after oral administration of [^{14}C]cholesterol in olive oil (Fig. 3B). The level of free [^{14}C]cholesterol was similar between groups (fig. S6B). SREBP-2 target gene expression was higher in global B/C-KO than in WT mice, both in fasted and refed states (Fig. 3, C and D), consistent with reduced ER cholesterol. Similar findings were observed in jejunal scrapings of I-B/C-KO mice (Fig. 3E). Protein levels of SREBP-2 targets in duodenal scrapings were correspondingly elevated in B/C-KO mice (Fig. 3F). We also observed up-regulation of *Srebf2* and its targets in jejunal scrapings in B/C-KO mice fed a high-cholesterol (HC) diet (1.25% cholesterol) (fig. S6C). Thus, even in the setting of increased dietary cholesterol, cholesterol synthesis was activated in B/C-KO mice. We did not observe changes in SREBP-2 pathway expression in the jejunum of global (fig. S6, D and E) or intestinal-specific (fig. S6, F and G) single-KO mice.

Plasma levels of ApoB48 after refeeding with a HC diet were not different between groups (Fig. 3, G and H), despite lower plasma cholesterol levels in I-B/C KO mice (Fig. 3I). Plasma triglyceride levels were also comparable (fig. S6H), consistent with our observation that fatty acid absorption was unaffected by loss of Asters (fig. S4, A to D). These findings suggest that Aster deficiency reduces cholesterol content in chylomicrons, but does not alter the process of chylomicron assembly and release per se. In support of this conclusion, there were fewer ^{14}C counts in the chylomicron fraction of plasma isolated 2 hours after oral gavage of [^{14}C]cholesterol in I-B/C KO mice versus control mice (Fig. 3J). To show that this decrease in cholesterol was due to a reduction in CE, we analyzed chylomicron composition in mice gavaged with d4-cholesterol and injected with Poloxamer-407. I-B/C KO mice had reduced d4-CE in chylomicrons (Fig. 3K) without changes in d4-cholesterol compared with floxed control mice (Fig. 3L). Unlabeled CE species were also lower whereas unlabeled cholesterol was not different (fig. S6, I and J).

Loss of Asters protects against diet-induced hypercholesterolemia

Next, we investigated whether intestinal Aster deficiency affected systemic cholesterol homeostasis. After 21 days of an HC diet we observed a modest decrease in the weight of global B/C-KO mice (Fig. 4A), accompanied by reduced fasting plasma cholesterol levels (Fig. 4B). Liver cholesterol levels were also reduced (Fig. 4C). We also found lower fasting plasma cholesterol levels in intestine-specific Aster-KO mice but no change in body weight (Fig. 4, D and E). FPLC fractionation of plasma revealed lower cholesterol levels in the VLDL/LDL and HDL fractions from Aster-KO mice (Fig. 4F). Plasma

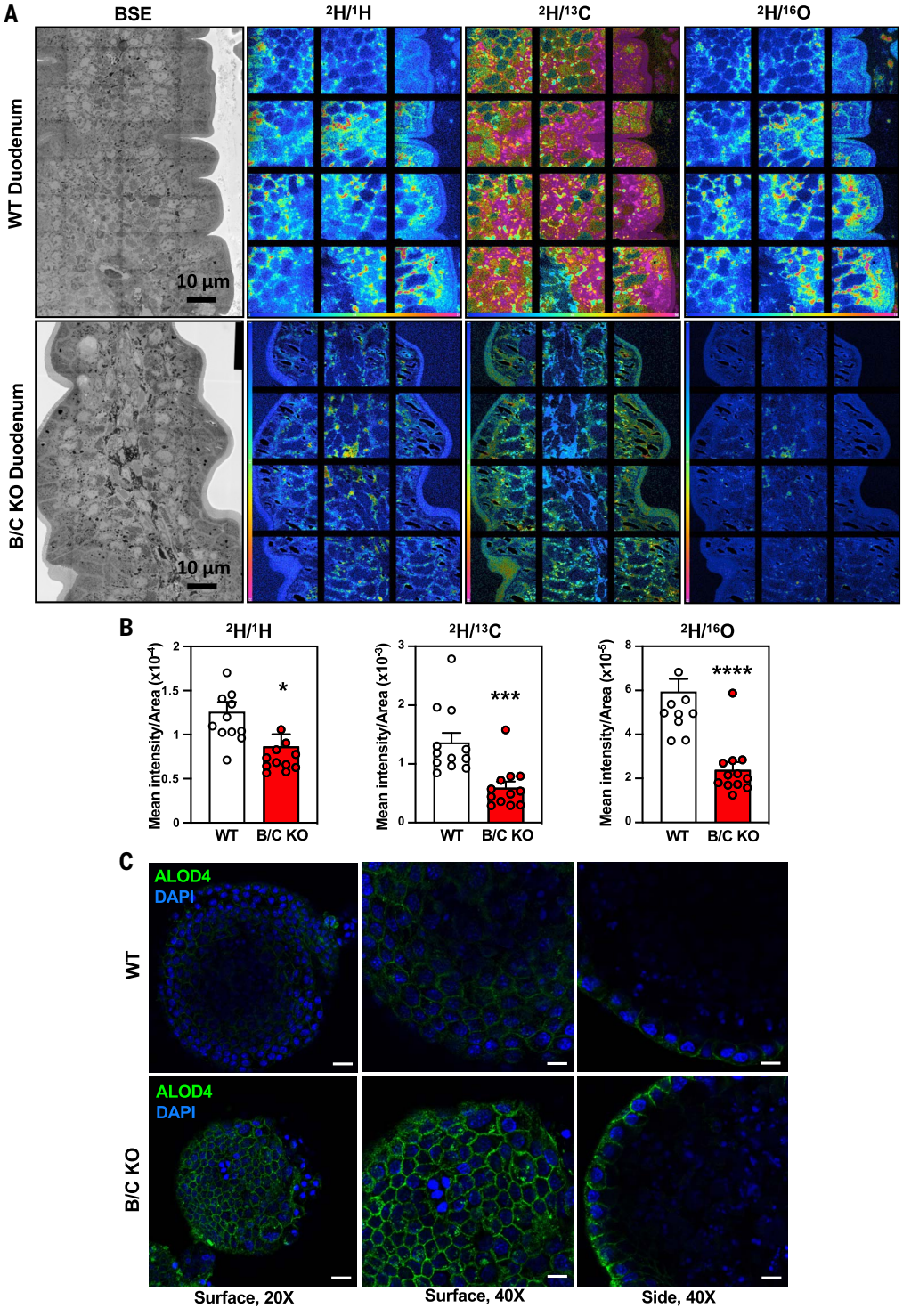


Fig. 2. Deletion of Asters reduces cholesterol internalization from the plasma membrane. (**A**) Backscattered electron images and NanoSIMS images of a mid-duodenum villus from WT and B/C KO mice. (**B**) Quantification of the $^{2}H^{-}$ secondary ion signal, normalized to the $^{1}H^{-}$, $^{13}C^{-}$, and $^{16}O^{-}$ signals. (**C**) ALOD4 imaging of murine enteroids from WT and B/C KO mice 2 hours after loading with MβCD cholesterol. For left panels, scale bar is 20 µm. For middle and right panels scale bar is 10 µm. Data are expressed as mean ± SEM. Statistical analysis: unpaired *t* test; **P* < 0.05, ****P* < 0.001, *****P* < 0.0001.

triglyceride levels were similar between groups (fig. S7, A and B). SREBP-2 pathway gene and protein expression (Fig. 4, G to I) was elevated in global B/C-KO and I-B/C-KO mice compared with controls. Thus, de novo cholesterol synthesis was activated in the KO intestine even in the setting of excess dietary cholesterol (Fig. 4, H and I). Lipidomic analyses revealed reduced CE in I-B/C-KO enterocytes on the HC diet (Fig. 4J), without changes in other lipid species (fig. S7C). We did not observe changes in body weight (fig. S7D) or lipid levels in Aster single-KO mice after 21 days of the HC diet (fig. S7, E and F), nor did we observe a change in SREBP-2 targets (fig. S7G).

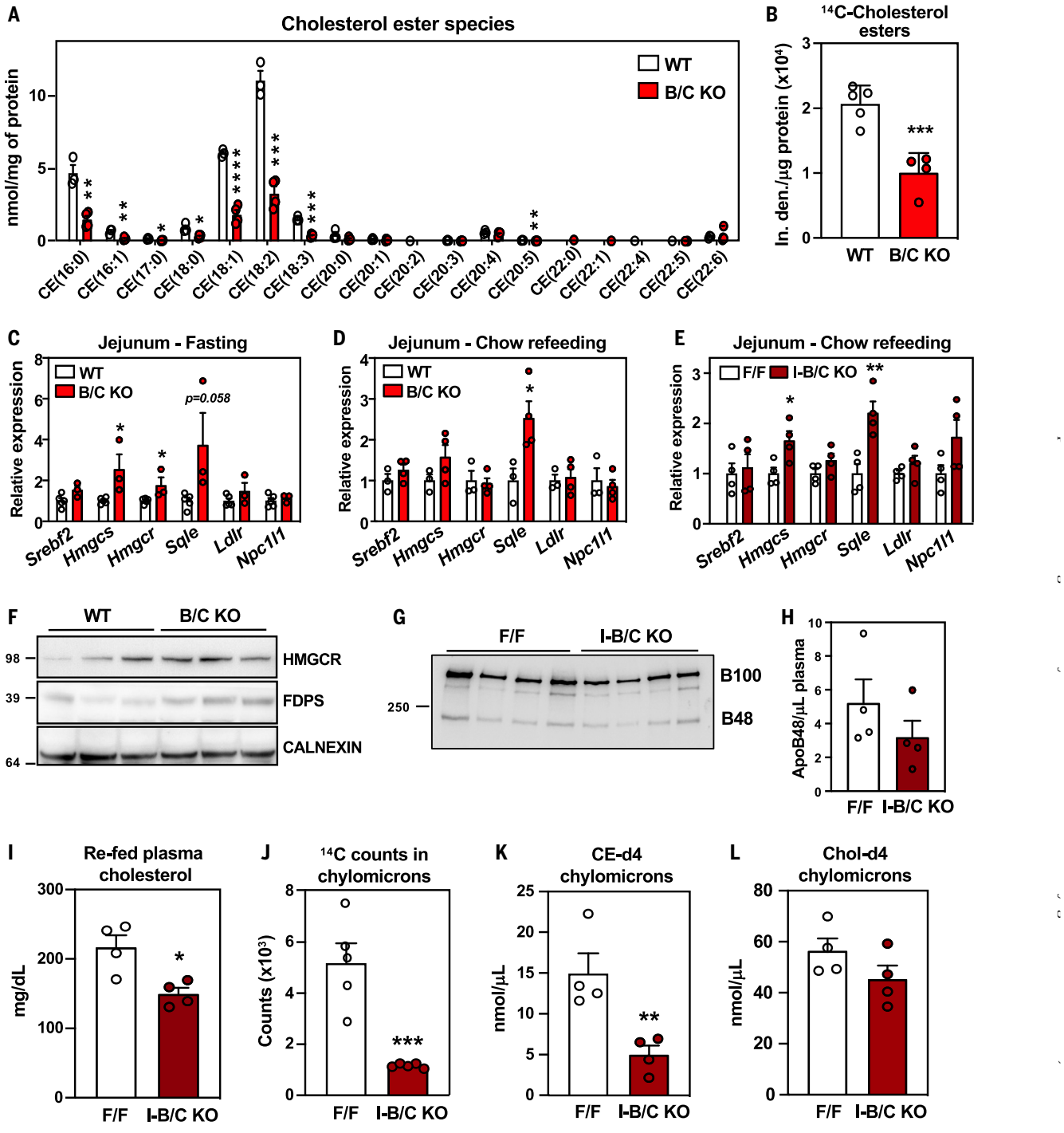


Fig. 3. Loss of Asters in intestine impairs cholesterol transfer to the ER. (**A**) Cholesterol ester quantification by mass spectrometry in proximal jejunum from WT (n = 3) and B/C KO (n = 4) 2 hours after refeeding a chow diet following a 10-hour fast. (**B**) Quantification of ^{14}C-labeled CE isolated from proximal jejunum scrapings of WT (n = 5) and B/C KO (n = 4) mice 2 hours after oral gavage of [^{14}C]cholesterol. (**C**) Gene expression from distal jejunum scrapings of WT (n = 5) and B/C KO (n = 3) mice after 4 hours of fasting. (**D**) Gene expression from distal jejunum scrapings of WT (n = 3) and B/C KO (n = 4) mice 2 hours after refeeding chow diet following a 10-hour fast. (**E**) Gene expression from distal jejunum scrapings of F/F (n = 4) and I-B/C KO (n = 4) mice 2 hours after refeeding chow diet following a 10-hour fast. (**F**) Western blot analysis of duodenum scrapings of mice described in (D). (**G**) Western blot analysis and quantification of plasma from F/F (n = 4) and I-B/C KO (n = 4) mice fed for 21 days a high cholesterol (1.25%) diet, after a 10-hour fast followed by 2 hours of refeeding a HC diet. (**H**) ApoB48 quantified by densitometry and normalized on the volume of plasma used for WB detection. (**I**) Plasma cholesterol of mice described in (G). (**J**) Quantification of ^{14}C-counts in chylomicrons isolated from plasma of F/F (n = 5) and I-B/C KO (n = 5) 3 hours after treatment with Poloxamer-407 and oral gavage of [^{14}C]cholesterol. (**K**) and (**L**) Quantification of deuterated (-d4) CE (K) and free cholesterol (L) in chylomicrons isolated from plasma of F/F (n = 4) and I-B/C KO (n = 4) 3.5 hours after treatment with Poloxamer-407 and oral gavage with cholesterol-d4. Data are expressed as mean ± SEM. Statistical analysis: unpaired t test, *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

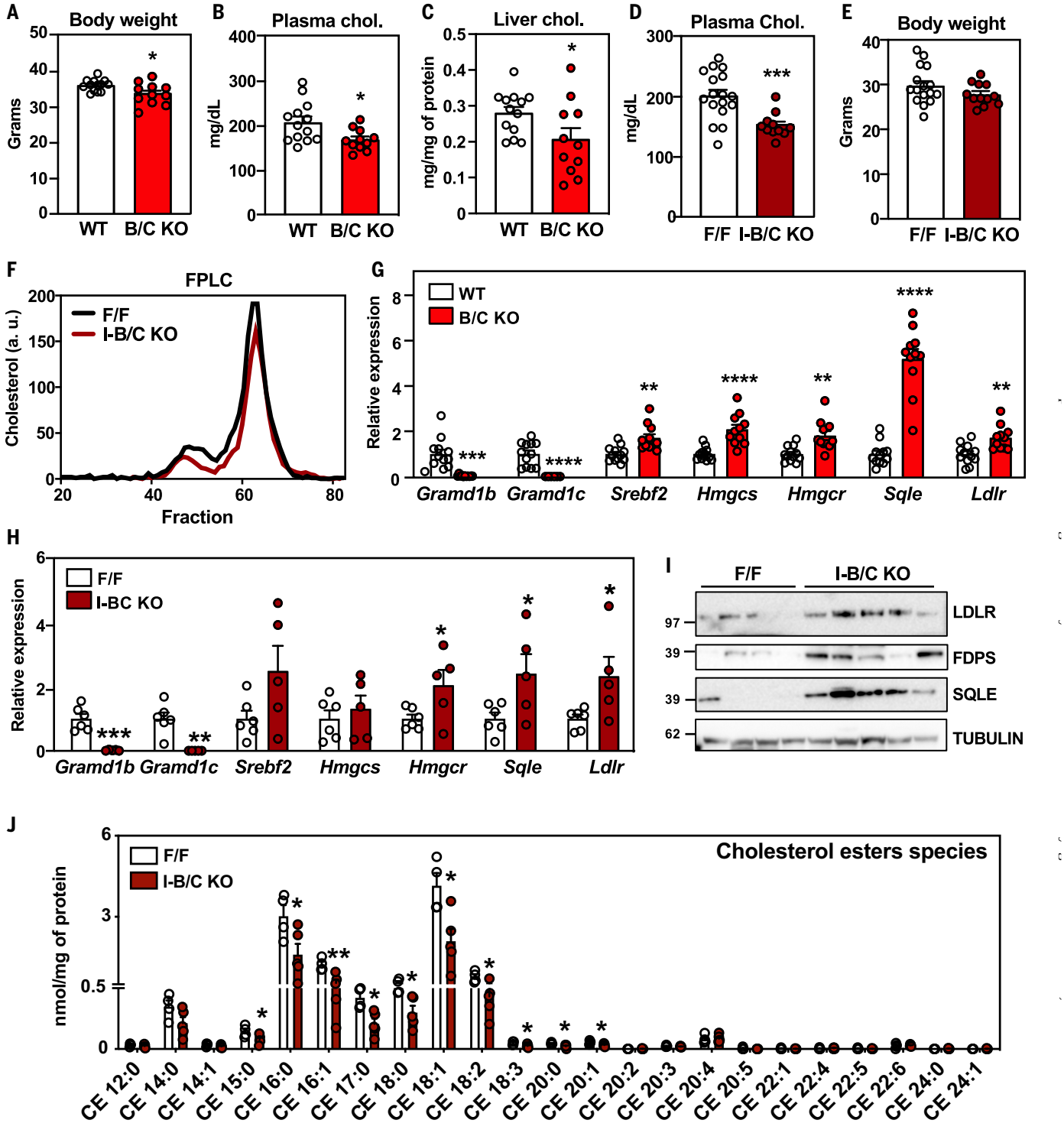


Fig. 4. Deletion of intestinal Asters protects from diet-induced hypercholesterolemia. (**A**) Body weight of WT (n = 13) and B/C KO (n = 11) male mice after 21 days of a Western diet plus 1.25% cholesterol (HC). (**B**) Plasma cholesterol levels after 4 hours fasting in mice described in (A). (**C**) Liver cholesterol in mice described in (A). (**D**) Plasma cholesterol levels after 4 hours fasting in F/F (n = 17) and I-B/C KO (n = 11) male mice after 21 days of the HC diet. (**E**) Body weight of mice described in (D). (**F**) FPLC of plasma from F/F and I-B/C KO mice fed for 21 days with HC diet and euthanized after overnight fasting followed by 2 hours of refeeding with the HC diet (pool of 3 to 5 mice per group). (**G**) Gene expression in distal jejunum scrapings from WT (n = 12) and B/C KO (n = 11) mice after 21 days of the HC diet, euthanized after 4 hours fasting. (**H**) Gene expression in distal jejunum scrapings from F/F (n = 6) and I-B/C KO (n = 5) male mice after 21 days of the HC diet, euthanized after 4 hours fasting. (**I**) Western blot analysis of duodenum scrapings of mice described in (H) (n = 4 to 5). (**J**) Lipidomic analysis of CE in proximal jejunum scrapings from mice described in (H) (n = 5 per group). Data are expressed as mean ± SEM. Statistical analysis: unpaired t test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

Ezetimibe binds to Aster-B and Aster-C

Because both Aster deficiency and EZ reduce cholesterol absorption, we tested whether Asters might bind EZ. Using competition assays for 22-NBD-cholesterol binding, we found that EZ bound to Aster-B and Aster-C (Fig. 5A) with moderate affinity but exhibited minimal binding to Aster-A and StARD1 (fig. S8A). Next, we solved the crystal structure, at 1.6 Å resolution, of the ASTER domain from Aster-C complexed to EZ. The overall structure of the Aster-C domain revealed a canonical curved seven-stranded beta-sheet that forms a cavity to accommodate EZ (Fig. 5B and table S3). The cavity is closed by a long carboxyl-terminal helix and two shorter helices. The electron density map and the simulated annealing composite omit map highlighted additional volume within the binding cavity that accommodates a glycerol molecule, and density for part of a PEG4000 molecule from the cryo-protectant and crystallization buffer (fig. S8, B and C, respectively). The electron density for the PEG is stronger at the end in proximity to the EZ ligand and becomes weaker at the other end, presumably due to disorder. Thus, the precise translational position of the PEG is uncertain. Modeling of ezetimibe-glucuronide—the active metabolite of EZ—in the pocket suggested potential capacity for binding as the glucuronide group is oriented toward an opening of the pocket (fig. S8D). Circular dichroism (CD) spectra of the Aster-A, -B, and -C domains revealed that Aster-B and -C were thermally stabilized by EZ binding, whereas Aster-A was not (fig. S8E). Cholesterol and U18666A were included as positive controls (*25*, *36*).

To investigate the molecular basis for selective binding of EZ to Aster-C, we compared the structures of Aster-C:EZ and Aster-A:25–hydroxycholesterol (Fig. 5C). Many of the residues involved in ligand interactions are conserved in Aster-A and Aster-C (fig. S8F). However, leucine 400 and phenylalanine 405 in the ASTER domain of Aster-A appeared to represent a steric hindrance to EZ binding (Fig. 5D), whereas alanine 357 and isoleucine 362 in the ASTER domain of Aster-C left ample space for EZ binding (Fig. 5D). The importance of the specific amino acids on EZ binding was interrogated by mutagenesis and evaluation of thermal stability by CD. The two residues that appeared to hinder EZ binding in Aster-A were changed to the residues found in Aster-C (L400A_F405I). Conversely, the residues that appeared to be important for accommodating EZ binding in Aster-C were changed to those in Aster-A (A357L_I362F), which potentially prevent binding. The Aster-A mutant L400A_F405I, as well as the single mutant L400A, reduced Aster-A thermal stability (fig. S8G), but the stability was increased by EZ. The thermal stability upon EZ binding was unchanged in the F405I mutant (fig. S8G), suggesting that L400A is crucial for binding selectivity. These results were corroborated by the Aster-C mutants, where A357L_I362F and A357L alone increased thermal stability of the protein whereas I362F had no effect. The thermal stability of the A357L Aster-C mutant did not change with EZ binding, implying a weak interaction between EZ and that mutant (fig. S8G).

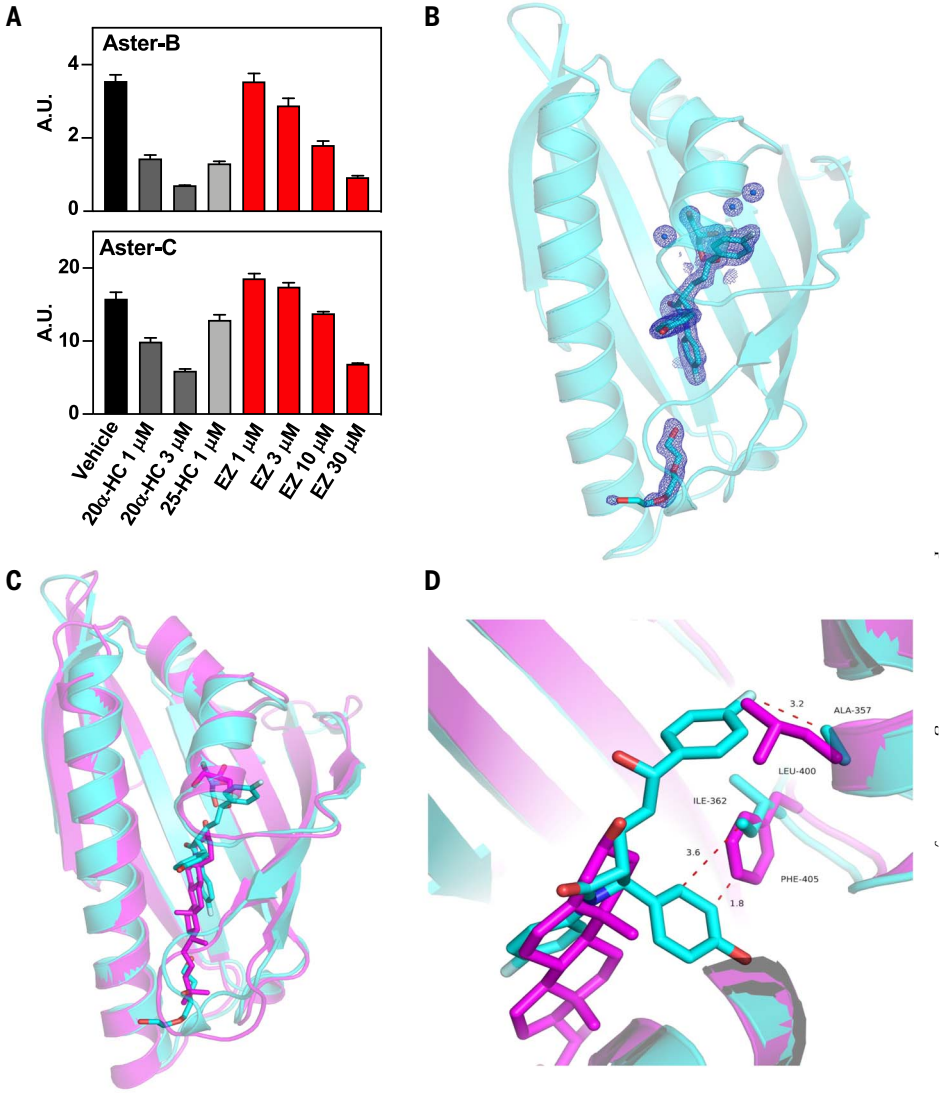


Fig. 5. Aster-B and Aster-C bind ezetimibe. (**A**) Competition assays for 22-NBD-cholesterol binding to purified ASTER-B and ASTER-C domains incubated in the presence of a vehicle, 20 α-HC, 25-HC, or EZ (1 to 30 mM). Data bars represent means ± SD. (**B**) Illustrated representation of the atomic structure at 1.6 Å resolution of the ASTER domain of Aster-C complexed to EZ. The EZ ligand in cyan sticks fits in a pocket created between the highly curved β-sheet, the second short helix, and the carboxyl-terminal helix. The cavity also accommodates glycerol (cyan sticks), water molecules (blue spheres), and part of a PEG 4000 molecule (cyan sticks). 2Fo-Fc electron density map of EZ, glycerol, part of PEG 4000, and water molecules in the binding cavity shown as blue mesh contoured at 1.2 σ. (**C**) Superposition of Aster-C:EZ (cyan) with the structure of Aster-A:25-HC (pdb ID 6GQF). The overall architecture of the domain is very similar but there are differences in the binding pocket and the opening loop. (**D**) Detail of the interaction between EZ and Aster-C involving residues ALA 357 and ILE 362, and superposition of the Aster-A:25-HC interaction. The superposition reveals how LEU 400 and PHE 405 from the Aster-A domain would clash with EZ, potentially preventing Aster-A from binding EZ.

Asters cooperate with NPC1L1 in intestinal cholesterol absorption

The ability of Aster-B and -C to bind EZ suggested that they may act in concert with NPC1L1 to promote cholesterol absorption. To test this hypothesis, we fed mice a moderate-cholesterol control diet or a moderate-cholesterol diet containing 0.01% EZ for 3 days (fig. S9A). As

expected, EZ decreased CE content and fractional cholesterol absorption by the fecal dual isotope method in WT mice (Fig. 6, A to C). Aster-B/C-KO and I-B/C-KO mice showed decreased CE content (Fig. 6, A and B) and fractional cholesterol absorption (Fig. 6C); EZ treatment led to a further reduction. Aster deficiency did not provide an additional decrease in CE or cholesterol absorption in EZ-treated mice (Fig. 6, A to C). However, gene expression and protein analyses revealed stronger activation of the SREBP-2 pathway in Aster-B/C-KO mice on the EZ diet compared with either Aster-B/C KO mice or EZ-treated WT mice (fig. S9B and Fig. 6D).

Cholesterol loading triggers Aster-B movement to the PM (Fig. 1B). EZ blocks the channel within NPC1L1 required for cholesterol deposition into the PM (*15*). Therefore, we theorized that EZ would attenuate Aster translocation to the PM by preventing the expansion of the accessible cholesterol pool. To test this idea, 3×HA-Aster-B mice were fed control or EZ diet for 3 days and gavaged with vehicle or EZ 30 min prior to a second gavage of cholesterol in corn oil. Small intestines were harvested 1 hour later. EZ completely prevented the recruitment of Aster-B to the brush border after cholesterol gavage (Fig. 6E). Next, we everted enteroids such that the apical PM was facing out (*37*) and administered mixed micelles (MM) containing cholesterol in the presence of vehicle or EZ. Immunofluorescence microscopy confirmed that recruitment of HA-Aster-B to the apical PM of enteroids was induced by cholesterol loading in MM but reduced by EZ (fig. S9C). Furthermore, we found decreased ALOD4 staining in enteroids derived from NPC1L1-KO mice loaded with MM cholesterol, indicating that NPC1L1 is required to saturate the PM with accessible cholesterol (Fig. 6F, upper panels). When enteroids were loaded with MβCD cholesterol, ALOD4 binding was not different between NPC1L1KO and WT enteroids (Fig. 6F, bottom panels). Thus, delivery of cholesterol in MβCD effectively bypasses the function of NPC1L1.

We then crossed 3×HA-Aster-B mice to NPC1L1 WT (HA-B-NPC1L1 WT) or NPC1L1 KO (HA-B-NPC1L1 KO) animals. In enteroids derived from these mice, NPC1L1 deletion reduced 3×HA-Aster-B recruitment to the PM after MM cholesterol loading but had no effect on MβCD cholesterol loading (fig. S9D). Analysis of HA-B-NPC1L1 WT and HA-B-NPC1L1 KO mice confirmed that genetic ablation of NPC1L1 abolished Aster-B movement to the brush border in response to cholesterol loading, mirroring the effects of EZ (Fig. 6G and fig. S9E). These results substantiate the hypothesis that EZ prevents Aster translocation to the brush border through NPC1L1 blockade. They further show that EZ prevents the saturation of cholesterol at the apical membrane by inhibiting NPC1L1 rather than Aster-B/C inhibition.

Next, we generated an McA-RH7777 CRL-1601 cell line stably expressing NPC1L1-EGFP and HA-Aster-B fusion proteins. Following cholesterol depletion, NPC1L1 partially localized to the PM, whereas Aster-B was confined to the ER (fig. S9F). MβCD-cholesterol loading caused internalization of NPC1L1-EGF and, concomitantly, the movement of HA-Aster-B to ER-PM contact sites. This contrasting behavior of NPC1L1 and Aster in response to cholesterol loading argues against a direct physical interaction between the two proteins during cholesterol import.

Pharmacological Aster inhibition reduces intestinal cholesterol uptake

Previous work from our laboratory identified a small molecule, AI-3d, that potently inhibits Aster-A, -B, and -C (*36*). We tested the ability of AI-3d to mimic the effects of Aster deficiency on cholesterol absorption. First, we pretreated enteroids with AI-3d and assessed accessible cholesterol with the ALOD4 probe. Treatment of WT enteroids with AI-3d and MβCD cholesterol led to an accumulation of accessible cholesterol, whereas AI-3d had no additional effect on ALOD4 binding in B/C-KO enteroids (Fig. 7A and fig. S10A). AI-3d generated a similar effect in human jejunal enteroids (Fig. 7B), indicating that AI-3d inhibits both human and murine Asters. We also compared the effects of AI-3d on WT and NPC1L1 KO enteroids. When enteroids were loaded with MM cholesterol, AI-3d treatment increased ALOD4 binding only in NPC1L1 WT cells (fig. S10B). However, when enteroids were loaded with MβCD-cholesterol, AI-3d treatment increased ALOD4 binding similarly in WT and NPC1L1 KO cells (fig. S10C). Thus, NPC1L1 deposits micellular cholesterol into the PM to expand the accessible pool and MβCD bypasses this function.

Finally, we assessed the impact of AI-3d on cholesterol transport to the ER. Consistent with our results in Aster-KO mice, AI-3d increased SREBP-2 target gene expression in human enteroids (Fig. 7C) and differentiated Caco-2 cells (fig. S10D). We further tested the effects of AI-3d on cholesterol absorption in vivo. C57BL/6J mice were pretreated with AI-3d or EZ for 1 hour, gavaged with [^{14}C]cholesterol, and given an intraperitoneal injection of Poloxamer-407. EZ treatment and pharmacologic inhibition of Asters both reduced dietary cholesterol absorption into the circulation (Fig. 7D).

Next, we evaluated the impact of Aster inhibition on mice fed a moderate-cholesterol diet (fig. S10E). AI-3d did not affect body weight or intestinal length (fig. S10, F to H), supporting the absence of toxicity. Lipidomic analysis revealed reduced CE in jejunal scrapings from NPC1L1 KO compared with WT mice. Aster inhibition by AI-3d lowered levels of jejunal CE similarly in WT and NPC1L1 KO, indicating that AI-3d does not act through NPC1L1 (Fig. 7E). Gene expression analysis showed a further induction of SREBP2 targets in NPC1L-KO mice treated with AI-3d (Fig. 7F).

Finally, we tested the effect of AI-3d on NPC1L1 recycling using McA-RH7777 CRL-1601 cells stably expressing an NPC1L1-EGFP fusion protein. Consistent with published data (*17*), we observed that NPC1L1-EGFP was present within a recycling compartment when both vehicle-treated and AI-3d treated cells were cultured in full serum (fig. S10I, left panels). In response to cholesterol depletion by MβCD, the signal in both cell types moved to the PM (fig. S10I, middle panels). Conversely, MβCD-cholesterol loading led to the relocation of NPC1L1-EGFP to endosomes (fig. S10I, right panels). These results indicate that AI-3d treatment does not interfere with the process of NPC1L1 recycling, further supporting the conclusion that AI-3d inhibits cholesterol absorption by targeting Asters.

Discussion

The involvement of NPC1L1 in facilitating the entry of cholesterol into enterocytes and the role of ACAT2 in cholesterol esterification are well documented and both proteins are recognized to be key players in the process of cholesterol absorption (*5*, *6*). However, how cholesterol that enters the cell through NPC1L1 reaches the ER for esterification and regulation of cholesterol synthesis has been a long-standing mystery. Here we solve that mystery by showing that Aster-B and -C link NPC1L1 to ACAT2 by facilitating nonvesicular cholesterol transport to the ER following uptake by NPC1L1 at the enterocyte PM. Combined deletion of Aster-B and -C impairs the movement of dietary cholesterol to ER in enterocytes, as evidenced by expansion of the accessible PM cholesterol pool, reduced CE formation, and activation of the SREBP-2 pathway for cholesterol synthesis. Physiologically, inhibition of Aster nonvesicular transport reduces cellular cholesterol stores and impairs the incorporation of CE into chylomicrons.

Our results suggest that cholesterol homeostasis in enterocytes requires both NPC1L1-mediated cholesterol deposition into the PM and subsequent Aster-mediated transport to the ER. Because NPC1L1 is the gatekeeper that controls the first step of cholesterol uptake, cholesterol absorption is reduced when NPC1L1 is blocked by EZ. In this context, Aster deletion does not further decrease absorption, because the apical PM cholesterol available for transfer is limited. At the same time, the observation that CE production in the ER is markedly impaired by Aster deletion even when NPC1L1 is present argues against a dominant role for NPC1L1 itself in cholesterol transport to the ER. Thus, NPC1L1 and Asters function in series to move cholesterol from the diet into circulation. EZ binds to Aster-B

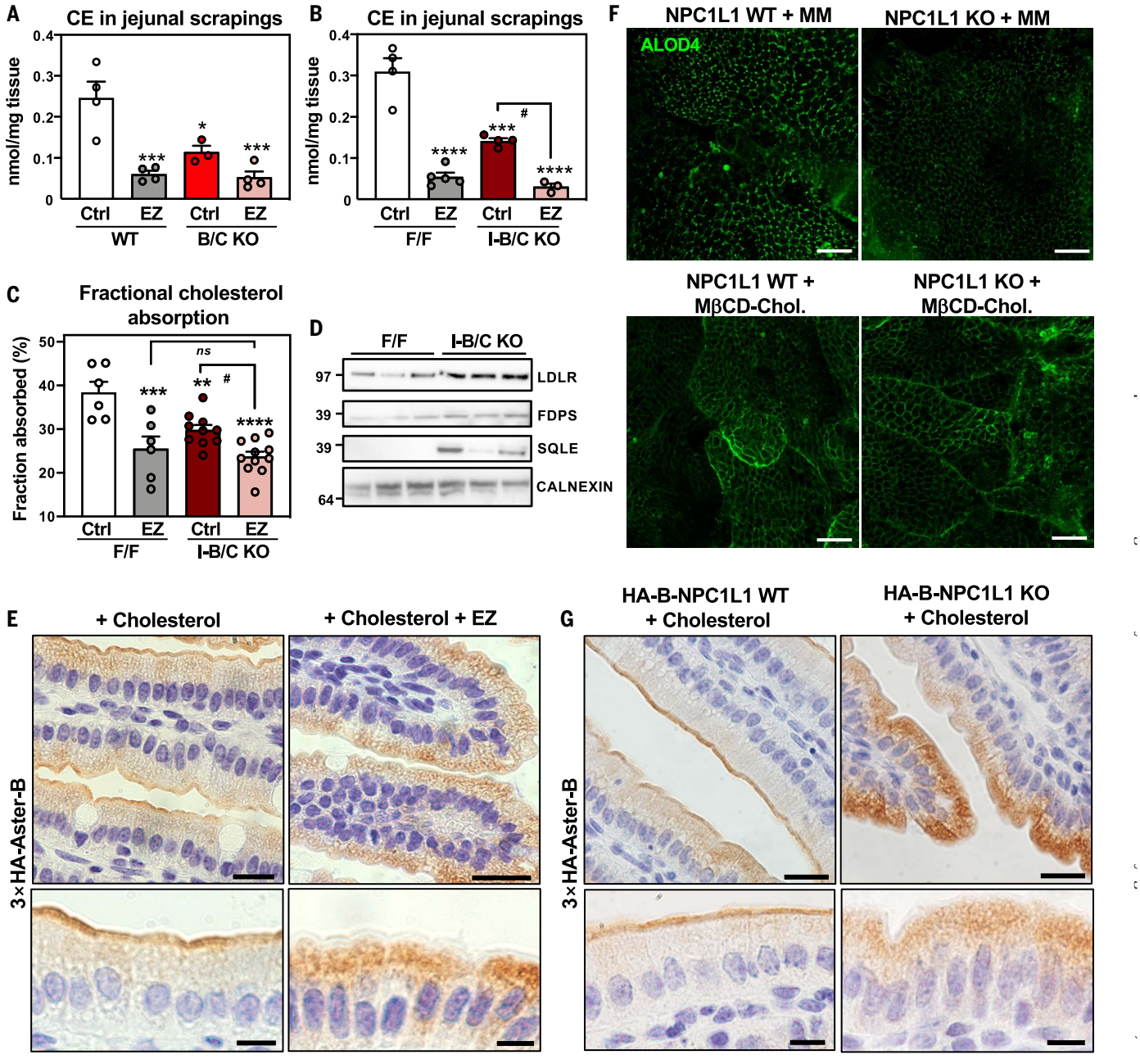


Fig. 6. NPC1L1 enriches accessible cholesterol at the brush border and promotes Aster recruitment. (**A**) CE quantification by mass spectrometry in scrapings from proximal jejunum of WT and B/C KO mice fed for 3 days with a control diet (ctrl) containing 0.08% cholesterol or a diet containing 0.08% cholesterol and 0.01% EZ and euthanized after 10 hours fasting followed by 2 hours refeeding with the same diets (n = 3 to 5 per group). (**B**) CE quantification by mass spectrometry in scrapings from proximal jejunum of F/F and I-B/C KO mice fed for 3 days with a control diet (ctrl) containing 0.08% cholesterol or a diet containing 0.08% cholesterol and 0.01% EZ) and euthanized 2 hours after refeeding (n = 3 to 5 per group). (**C**) Fractional absorption of cholesterol measured by the fecal dual isotope method in F/F and I-B/C KO mice fed for 3 days with a control diet (ctrl) containing 0.08% cholesterol or a diet containing 0.08% cholesterol and 0.01% EZ. (**D**) Western blot analysis of duodenum scrapings from F/F and I-B/C KO mice fed for 3 days with EZ diet (n = 3 per group). Samples were run on the same gel used for Western blot analysis reported in fig. S3F; therefore the loading control (Calnexin) is the same. (**E**) Immunohistochemistry of HA-Aster-B in small intestines from 3×HA-Aster-B mice after oral administration of vehicle or EZ and a gastric gavage with cholesterol in corn oil. (**F**) ALOD4 imaging of murine enteroids from WT and NPC1L1 KO mice after loading with cholesterol in mixed micelles or with MβCD-cholesterol. Scale bar is 40 μm. (**G**) Immunohistochemistry of HA-Aster-B in small intestines from 3×HA-Aster-B mice crossed to NPC1L1 WT or NPC1L1 KO mice after a gastric gavage with cholesterol in corn oil. For (E) and (G) upper and lower panels, scale bar is 20 and 10 μm, respectively. Data are expressed as mean ± SEM. 2-way ANOVA with Tukey's multiple comparisons test. Data are expressed as mean ± SEM. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$ versus WT Ctrl, $^{\#}P < 0.05$ versus I-B/C KO Ctrl.

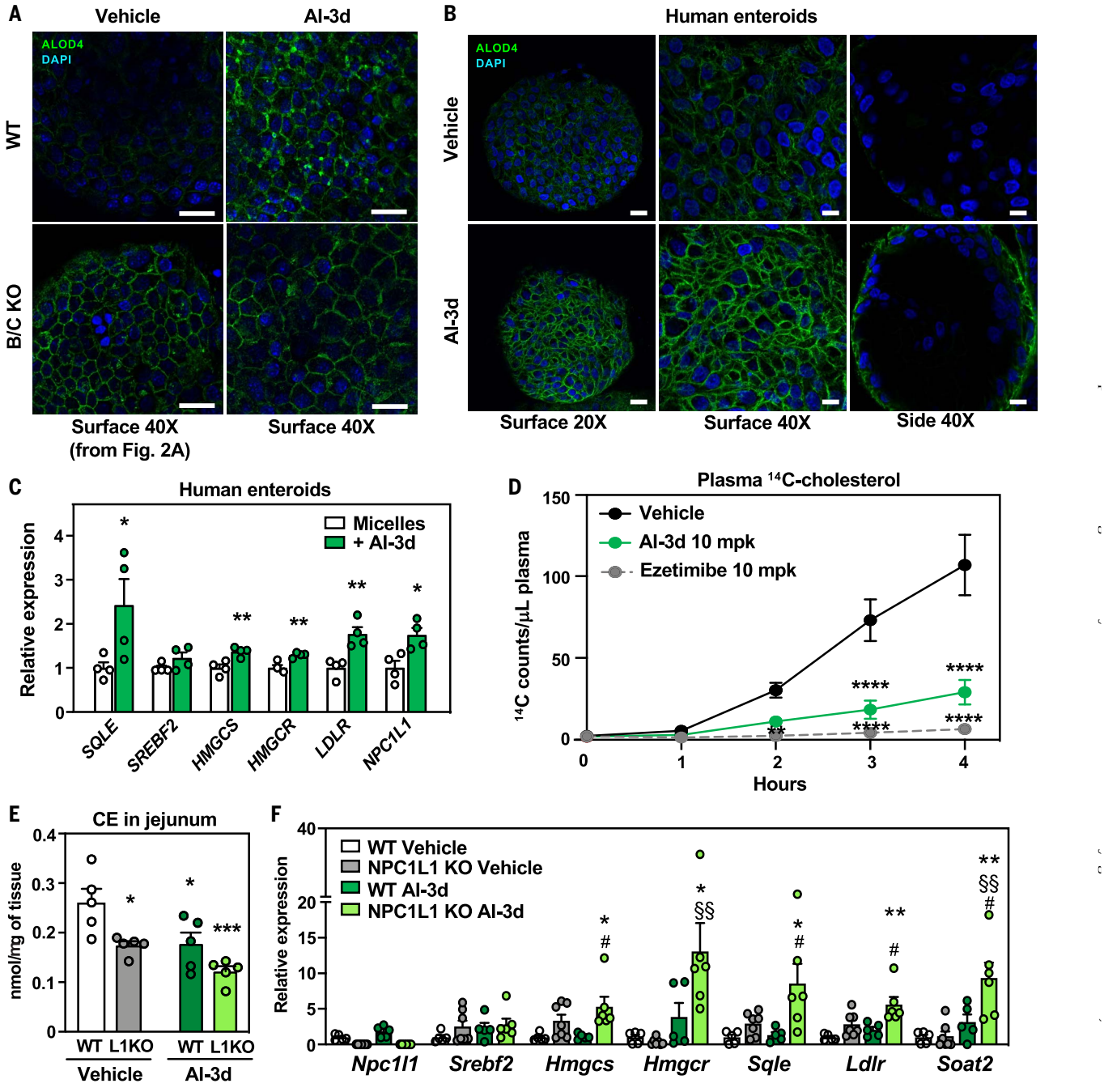


Fig. 7. Pharmacological inhibition of Asters reduces intestinal cholesterol uptake. (**A**) ALOD4 staining (green) of PM accessible cholesterol in enteroids from WT and B/C KO mice loaded with MβCD cholesterol and treated with vehicle or Aster inhibitor Al-3d. Results presented in this panel and in Fig. 2C came from one experiment, where WT vehicle-treated enteroids served as control for both B/C KO vehicle-treated enteroids and for Al-3d vehicle-treated enteroids. Scale bar is 20 µm. (**B**) ALOD4 staining (green) of PM accessible cholesterol in human enteroids treated with vehicle or Al-3d. For left panels, scale bar is 20 µm. For middle and right panels, scale bar is 10 µm. (**C**) Gene expression of SREBP-2 and target genes in human enteroids differentiated on transwell plates and loaded with cholesterol in mixed micelles in the presence of vehicle or Al-3d. (**D**) Kinetics of radioactivity in plasma of female mice that were administered vehicle (n = 6), Al-3d (n = 5), or EZ (n = 6) in corn oil. After 1 hour, the mice were given a gastric gavage of olive oil containing [^{14}C]cholesterol. (**E**) CE quantification by mass spectrometry in scrapings from proximal jejunum of NPC1L1 WT and NPC1L1 KO mice fed for 3 days with a control diet (ctrl) containing 0.08% cholesterol and treated with 3 doses of vehicle (left) or 10 mg/kg Al-3d (right) and euthanized after 10 hours of fasting followed by 2 hours refeeding with the same diet (n = 5 per group). (**F**) Gene expression analysis of distal jejunum scrapings from mice described in (E). Statistical analysis for (C), unpaired t test; for s, (D), (E), (F), 2-way ANOVA with Tukey's multiple comparisons test. Data are expressed as mean ± SEM. For (C), $^{*}P < 0.05$, $^{**}P < 0.01$ versus micelles. For (D) $^{**}P < 0.01$, $^{****}P < 0.0001$ versus Vehicle. For s (D), (E) $^{*}P < 0.05$, $^{**}P < 0.01$ versus WT Vehicle, $^{\#}P < 0.05$ versus WT Al-3d, $^{\S\S}P < 0.01$ versus NPC1L1 KO Vehicle.

and -C with moderate affinity, raising the possibility that inhibition of Aster function could contribute to EZ's capacity to inhibit cholesterol absorption. However, EZ binding to Aster was not sufficient to block nonvesicular cholesterol transport in vivo in our studies, as Aster deletion further reduced ER cholesterol delivery in the presence of EZ.

The recruitment of Asters to the PM depends on the cholesterol and phosphatidylserine content in the inner leaflet (*25*, *26*, *33*). Our results support the notion that Aster-mediated nonvesicular transport relies on NPC1L1 to deposit cholesterol into the apical enterocyte PM, rather than on a direct physical interaction with NPC1L1. In support of this concept, micelle-derived cholesterol cannot contribute to the expansion of the accessible pool, and Aster proteins do not translocate to the enterocyte PM when NPC1L1 is deleted or inhibited by EZ. However, when NPC1L1 is bypassed by delivering cholesterol directly to the PM with MβCD, Aster is recruited normally.

Previous studies have demonstrated that NPC1L1 is internalized when enterocytes are exposed to dietary cholesterol (*38*). This observation has led to the suggestion that vesicular transport by NPC1L1-containing vesicles is important for cholesterol delivery to the ER (*21*, *24*, *38*). However, formal tests of the relative contribution of vesicular and nonvesicular cholesterol transport mechanisms downstream of NPC1L1 have been lacking. Our data indicate that the primary function of NPC1L1 is to deposit dietary cholesterol into the enterocyte PM. NPC1L1 is a homolog of the Niemann-Pick C1 (NPC1) protein that mediates cholesterol export from lysosomes following endocytosis of LDL cholesterol (*39*). Molecular studies have established that NPC1 acts to enrich cholesterol in the lysosomal membrane (*40*, *41*), and that such enrichment allows for the subsequent transfer of cholesterol to PM by cytoplasmic nonvesicular transfer proteins (*42*). Our results suggest that the structurally related NPC1 and NPC1L1 perform analogous functions: both enrich membrane cholesterol to facilitate the recruitment of nonvesicular transporters. We propose that the sterol-regulated endocytosis of NPC1L1 could represent a feedback mechanism to limit cholesterol absorption, rather than a major mechanism for vesicular cholesterol transfer to the ER. However, our data do not exclude the participation of other vesicular and nonvesicular pathways in the intracellular movement of cholesterol in enterocytes. Nor do they exclude the possibility that Aster proteins retrieve cholesterol, directly or indirectly, from NPC1L1-rich endosomes.

Intestinal nonvesicular cholesterol transport could be a target for treating diet-induced hypercholesterolemia. First, Aster-B/C knockout mice have low plasma cholesterol levels when fed a Western diet enriched in cholesterol. Second, we showed that the Aster pathway can be targeted pharmacologically to reduce cholesterol absorption. The inhibitor AI-3d, which had been shown previously (*36*) to inhibit Aster-mediated nonvesicular transport in vitro, reduced cholesterol transport to the ER and expanded the pool of accessible cholesterol at the PM of intestinal enteroids. Finally, treatment of mice with AI-3d diminished cholesterol absorption. These findings identify the Aster pathway as a potentially attractive pathway for limiting intestinal cholesterol absorption and reducing plasma cholesterol.

Materials and methods summary

Detailed materials and methods can be found in the supplementary materials (*43*), including generation of murine models, experimental conditions for in vivo studies, histology and immunohistochemistry, cholesterol absorption assays, lipid measurements, chylomicron isolation, isolation of intestinal epithelial cells, protein isolation and immunoblot assays, RNA extraction and gene expression analysis, in vivo back-scattered electron (BSE) microscopy and NanoSIMS, isolation of murine crypts, culture and expansion of enteroids (basolateral-out or apical-out), culture of human intestinal epithelial cells on transwell membranes for gene expression, immunofluorescence microscopy, purification and fluorophore conjugation of ALOD4, immortalized cell cultures (Caco-2 and McArdle RH7777), protein expression and purification, crystallization and x-ray structure determination, and circular dichroism.

REFERENCES AND NOTES

1. P. Xie *et al*., Genetic demonstration of intestinal NPC1L1 as a major determinant of hepatic cholesterol and blood atherogenic lipoprotein levels. *Atherosclerosis* **237**, 609–617 (2014). doi: 10.1016/j.atherosclerosis.2014.09.036; pmid: 25463095
2. C. M. Mansbach, S. A. Siddiqi, The biogenesis of chylomicrons. *Annu. Rev. Physiol.* **72**, 315–333 (2010). doi: 10.1146/annurev-physiol-021909-135801; pmid: 20148678
3. K. K. Buhman *et al*., Resistance to diet-induced hypercholesterolemia and gallstone formation in ACAT2-deficient mice. *Nat. Med.* **6**, 1341–1347 (2000). doi: 10.1038/82153; pmid: 11100118
4. T. M. Nguyen, J. K. Sawyer, K. L. Kelley, M. A. Davis, L. L. Rudel, Cholesterol esterification by ACAT2 is essential for efficient intestinal cholesterol absorption: Evidence from thoracic lymph duct cannulation. *J. Lipid Res.* **53**, 95–104 (2012). doi: 10.1194/jlr.M018820; pmid: 22045928
5. J. Zhang *et al*., Tissue-specific knockouts of ACAT2 reveal that intestinal depletion is sufficient to prevent diet-induced cholesterol accumulation in the liver and blood. *J. Lipid Res.* **53**, 1144–1152 (2012). doi: 10.1194/jlr.M024356; pmid: 22460046
6. S. W. Altmann *et al*., Niemann-Pick C1 Like 1 protein is critical for intestinal cholesterol absorption. *Science* **303**, 1201–1204 (2004). doi: 10.1126/science.1093131; pmid: 14976318
7. H. R. Davis Jr. *et al*., Niemann-Pick C1 Like 1 (NPC1L1) is the intestinal phytosterol and cholesterol transporter and a key modulator of whole-body cholesterol homeostasis. *J. Biol. Chem.* **279**, 33586–33592 (2004). doi: 10.1074/jbc.M405817200; pmid: 15173162
8. M. Garcia-Calvo *et al*., The target of ezetimibe is Niemann-Pick C1-Like 1 (NPC1L1). *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8132–8137 (2005). doi: 10.1073/pnas.0500269102; pmid: 15928087
9. H. R. Davis Jr., K. K. Pula, K. B. Alton, R. E. Burrier, R. W. Watkins, The synergistic hypocholesterolemic activity of the potent cholesterol absorption inhibitor, ezetimibe, in combination with 3-hydroxy-3-methylglutaryl coenzyme a reductase inhibitors in dogs. *Metabolism* **50**, 1234–1241 (2001). doi: 10.1053/meta.2001.26737; pmid: 11586500
10. C. P. Cannon *et al*., Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. *N. Engl. J. Med.* **372**, 2387–2397 (2015). doi: 10.1056/NEJMoa1410489; pmid: 26039521
11. J. W. Clader, The discovery of ezetimibe: A view from outside the receptor. *J. Med. Chem.* **47**, 1–9 (2004). doi: 10.1021/jm030283g; pmid: 14695813
12. H. R. Davis Jr., D. S. Compton, L. Hoos, G. Tetzloff, Ezetimibe, a potent cholesterol absorption inhibitor, inhibits the development of atherosclerosis in ApoE knockout mice. *Arterioscler. Thromb. Vasc. Biol.* **21**, 2032–2038 (2001). doi: 10.1161/hq1201.100260; pmid: 11742881
13. R. G. Lee *et al*., Plasma cholesteryl esters provided by lecithin: cholesterol acyltransferase and acyl-coenzyme a:cholesterol acyltransferase 2 have opposite atherosclerotic potential. *Circ. Res.* **95**, 998–1004 (2004). doi: 10.1161/01.RES.0000147558.15554.67; pmid: 15486318
14. M. Hu *et al*., Structural insights into the mechanism of human NPC1L1-mediated cholesterol uptake. *Sci. Adv.* **7**, eabg3188 (2021). doi: 10.1126/sciadv.abg3188; pmid: 34272236
15. C.-S. Huang *et al*., Cryo-EM structures of NPC1L1 reveal mechanisms of cholesterol transport and ezetimibe inhibition. *Sci. Adv.* **6**, eabb1989 (2020). doi: 10.1126/sciadv.abb1989; pmid: 34407950
16. T. Long, Y. Liu, Y. Qin, R. A. DeBose-Boyd, X. Li, Structures of dimeric human NPC1L1 provide insight into mechanisms for cholesterol absorption. *Sci. Adv.* **7**, eabh3997 (2021). doi: 10.1126/sciadv.abh3997; pmid: 34407950
17. L. Ge *et al*., The cholesterol absorption inhibitor ezetimibe acts by blocking the sterol-induced internalization of NPC1L1. *Cell Metab.* **7**, 508–519 (2008). doi: 10.1016/j.cmet.2008.04.001; pmid: 18522832
18. A. B. Weinglass *et al*., Extracellular loop C of NPC1L1 is important for binding to ezetimibe. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11140–11145 (2008). doi: 10.1073/pnas.0800936105; pmid: 18682566
19. L. Ge *et al*., Flotillins play an essential role in Niemann-Pick C1-like 1-mediated cholesterol uptake. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 551–556 (2011). doi: 10.1073/pnas.1014434108; pmid: 21187433
20. L.-J. Wang *et al*., Molecular characterization of the NPC1L1 variants identified from cholesterol low absorbers. *J. Biol. Chem.* **286**, 7397–7408 (2011). doi: 10.1074/jbc.M110.178368; pmid: 21189420
21. P.-S. Li *et al*., The clathrin adaptor Numb regulates intestinal cholesterol absorption through dynamic interaction with NPC1L1. *Nat. Med.* **20**, 80–86 (2014). doi: 10.1038/nm.3417; pmid: 24336247
22. B.-B. Chu *et al*., Requirement of myosin Vb.Rab11a.Rab11-FIP2 complex in cholesterol-regulated translocation of NPC1L1 to the cell surface. *J. Biol. Chem.* **284**, 22481–22490 (2009). doi: 10.1074/jbc.M109.034355; pmid: 19542231
23. T. A. Johnson, S. R. Pfeffer, Ezetimibe-sensitive cholesterol uptake by NPC1L1 protein does not require endocytosis. *Mol. Biol. Cell* **27**, 1845–1852 (2016). doi: 10.1091/mbc.e16-03-0154; pmid: 27075173
24. Y.-Y. Zhang *et al*., A *LIMA1* variant promotes low plasma LDL cholesterol and decreases intestinal cholesterol absorption. *Science* **360**, 1087–1092 (2018). doi: 10.1126/science.aao6575; pmid: 29880681
25. J. Sandhu *et al*., Aster Proteins Facilitate Nonvesicular Plasma Membrane to ER Cholesterol Transport in Mammalian Cells. *Cell* **175**, 514–529.e20 (2018). doi: 10.1016/j.cell.2018.08.033; pmid: 30220461
26. T. Naito *et al*., Movement of accessible plasma membrane cholesterol by the GRAMD1 lipid transfer protein complex. *eLife* **8**, e51401 (2019). doi: 10.7554/eLife.51401; pmid: 31724953
27. M. Besprozvannaya *et al*., GRAM domain proteins specialize functionally distinct ER-PM contact sites in human cells. *eLife* **7**, e31019 (2018). doi: 10.7554/eLife.31019; pmid: 29469807
28. X. Xiao *et al*., Hepatic nonvesicular cholesterol transport is critical for systemic lipid homeostasis. *Nat. Metab.* **5**, 165–181 (2023). doi: 10.1038/s42255-022-00722-6; pmid: 36646756
29. B. Wang, P. Tontonoz, Liver X receptors in lipid signalling and membrane homeostasis. *Nat. Rev. Endocrinol.* **14**, 452–463 (2018). doi: 10.1038/s41574-018-0037-x; pmid: 29904174
30. L. Yu *et al*., Expression of ABCG5 and ABCG8 is required for regulation of biliary cholesterol secretion. *J. Biol. Chem.* **280**, 8742–8747 (2005). doi: 10.1074/jbc.M411080200; pmid: 15611112
31. H. Jiang *et al*., High-resolution imaging of dietary lipids in cells and tissues by NanoSIMS analysis. *J. Lipid Res.* **55**, 2156–2166 (2014). doi: 10.1194/jlr.M053363; pmid: 25143463

32. C. He *et al*., NanoSIMS Analysis of Intravascular Lipolysis and Lipid Movement across Capillaries and into Cardiomyocytes. *Cell Metab.* **27**, 1055–1066.e3 (2018). doi: 10.1016/j.cmet.2018.03.017; pmid: 29719224
33. A. Ferrari *et al*., Aster Proteins Regulate the Accessible Cholesterol Pool in the Plasma Membrane. *Mol. Cell. Biol.* **40**, e00255–e00220 (2020). doi: 10.1128/MCB.00255-20; pmid: 32719109
34. A. Das, M. S. Brown, D. D. Anderson, J. L. Goldstein, A. Radhakrishnan, Three pools of plasma membrane cholesterol and their relation to cholesterol homeostasis. *eLife* **40**, e02882 (2014). doi: 10.1128/MCB.00255-20; pmid: 32719109
35. A. Gay, D. Rye, A. Radhakrishnan, Switch-like responses of two cholesterol sensors do not require protein oligomerization in membranes. *Biophys. J.* **108**, 1459–1469 (2015). doi: 10.1016/j.bpj.2015.02.008; pmid: 25809258
36. X. Xiao *et al*., Selective Aster inhibitors distinguish vesicular and nonvesicular sterol transport mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024149118 (2021). doi: 10.1073/pnas.2024149118; pmid: 33376205
37. J. Y. Co *et al*., Controlling Epithelial Polarity: A Human Enteroid Model for Host-Pathogen Interactions. *Cell Rep.* **26**, 2509–2520.e4 (2019). doi: 10.1016/j.celrep.2019.01.108; pmid: 30811997
38. C. Xie *et al*., Ezetimibe blocks the internalization of NPC1L1 and cholesterol in mouse small intestine. *J. Lipid Res.* **53**, 2092–2101 (2012). doi: 10.1194/jlr.M027359; pmid: 22811412
39. E. D. Carstea *et al*., Niemann-Pick C1 disease gene: Homology to mediators of cholesterol homeostasis. *Science* **277**, 228–231 (1997). doi: 10.1126/science.277.5323.228; pmid: 9211849
40. H. J. Kwon *et al*., Structure of N-terminal domain of NPC1 reveals distinct subdomains for binding and transfer of cholesterol. *Cell* **137**, 1213–1224 (2009). doi: 10.1016/j.cell.2009.03.049; pmid: 19563754
41. H. Qian *et al*., Structural Basis of Low-pH-Dependent Lysosomal Cholesterol Egress by NPC1 and NPC2. *Cell* **182**, 98–111.e18 (2020). doi: 10.1016/j.cell.2020.05.020; pmid: 32544384
42. H. Wang *et al*., ORP2 Delivers Cholesterol to the Plasma Membrane in Exchange for Phosphatidylinositol 4, 5-Bisphosphate ($PI(4,5)P_2$). *Mol. Cell* **73**, 458–473.e7 (2019). doi: 10.1016/j.molcel.2018.11.014; pmid: 30581148
43. Materials and methods are available as supplementary materials.

ACKNOWLEDGMENTS

We thank L. Fong, P. Kim, and E. Tokhtaeva for the technical assistance. We also thank members of the Tontonoz, Tarling-Vallim, Edwards, Villanueva, Young, and Bensinger labs at UCLA for useful advice and discussions and for sharing reagents and resources. We thank G. Su and the UCLA Lipidomics and Y. Li and the UCLA Translational Pathology Core Laboratory. pALOD4 was a gift from Arun Radhakrishnan (Addgene plasmids # 111026). We thank B.-L. Song at Wuhan University for the NPC1L1-EGFP construct and NPC1L1 KO mice. We thank T. Darwish for supporting the production of [2H]cholesterol. **Funding:** This work was supported by the following: National Institutes of Health grant R01 DK126779 (to P.T.); National Institutes of Health grant P01 HL146358 (to S.G.Y.); Transatlantic Network of Excellence, Leducq Foundation, 19CDV04; American Diabetes Association Postdoctoral fellowship 1-19-PDF-043-RA and Ermenegildo Zegna Founder's Scholarship 2017 (to A.F.); CDI Junior Faculty Career Development Award (CD-JFCD-07012019), CDI NIH K12 Junior Faculty Career Development Grant (CDI-K12-07012023), and Today's and Tomorrow's Children Fund Bridge Grant (CDI-TTCF-07012022) (to E.W.); UCSD-UCLA Diabetes Research Center (DK063491) (to E.W. and P.T.); American Heart Association Postdoctoral Fellowship 18POST34030388 to (X.X.); American Heart Association Postdoctoral Fellowship 903306 to (J.P.K.); the Damon Runyon Cancer Research Fellowship (DRG-2424-21) to (Y.G.); National Institutes of Health grant T32 DK007180 to (A.N.); and the National Collaborative Research Infrastructure Strategy (NCRIS), an Australian Government initiative, to the National Deuteration Facility in Australia. **Author contributions:** Conceptualization: A.F., E.W., and P.T. Methodology: A.F., E.W., X.X., J.P.K., S.D.L., A.N., T.W., L.F., M.G.M., R.A.R., L.C., A.B., K.W. Investigation: A.F., E.W., X.X., J.P.K., B.R.A., J.J.M., T.W., K.C., J.S., M.J.T., L.B., A.N., Y.K., Y.G., P.M., and W.S. Visualization: A.F., E.W., and P.T. Funding acquisition: S.G.Y. and P.T. Project administration: P.T. Supervision: M.E.J., H.J., J.W.R.S., S.G.Y., and P.T. Writing – original draft: A.F., E.W., and P.T. Writing – review and editing: A.F., E.W., M.E.J., H.J., J.W.R.S., S.G.Y., P.T. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** Source data for all figures are provided with the paper. Sequencing data have been deposited to GEO (GSE206780). The crystal structure has been submitted to the PDB database, PDB ID code 8AXW. All unique biological materials used are readily available from the authors or from standard commercial sources. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adf0966
Summary
Materials and Methods
Figs. S1 to S10
Tables S1 to S3
References (*44*–*66*)

Submitted 28 September 2022; resubmitted 3 July 2023
Accepted 27 September 2023
10.1126/science.adf0966

RESEARCH ARTICLE

PHYSICS

Quantum control of trapped polyatomic molecules for eEDM searches

Loïc Anderegg[1,2]*, Nathaniel B. Vilas[1,2], Christian Hallas[1,2], Paige Robichaud[1,2], Arian Jadbabaie[3], John M. Doyle[1,2], Nicholas R. Hutzler[3]*

Ultracold polyatomic molecules are promising candidates for experiments in quantum science and precision searches for physics beyond the Standard Model. A key requirement is the ability to achieve full quantum control over the internal structure of the molecules. In this work, we established coherent control of individual quantum states in calcium monohydroxide (CaOH) and demonstrated a method for searching for the electron electric dipole moment (eEDM). Optically trapped, ultracold CaOH molecules were prepared in a single quantum state, polarized in an electric field, and coherently transferred into an eEDM-sensitive state where an electron spin precession measurement was performed. To extend the coherence time, we used eEDM-sensitive states with tunable, near-zero magnetic field sensitivity. Our results establish a path for eEDM searches with trapped polyatomic molecules.

The rich structure of polyatomic molecules makes them an appealing platform for experiments in quantum science (*1*–*4*), ultracold chemistry (*5*), and precision measurements (*6*–*10*). Key to this structure is the presence of near-degenerate states of opposite parity, which allow the molecules to be easily polarized in the laboratory frame with the application of a small electric field. Such states are generic among polyatomic molecules, but rare in diatomics, and may be useful for applications such as analog simulation of quantum magnetism models (*1*, *2*) or for realizing switchable interactions and long-lived qubit states for quantum computing (*4*). Additionally, the parity-doublet states in trapped polyatomic molecules are expected to be an invaluable tool for systematic error rejection in precision measurements of physics beyond the Standard Model (BSM) (*6*). To date, several species of polyatomic molecules have been laser cooled and/or trapped at ultracold temperatures (*11*–*17*).

One powerful avenue for tabletop BSM searches is probing for the electron electric dipole moment (eEDM) (*18*–*22*), d_e, which violates time-reversal (T) symmetry and is predicted by many BSM theories to be orders-of-magnitude larger than the Standard Model prediction (*19*, *20*). Present state-of-the-art eEDM experiments are broadly sensitive to T-violating physics at energies much greater than 1 TeV (*23*–*28*). All such experiments use Ramsey spectroscopy to measure an energy shift caused by the interaction of the electron with the large electric field that is present inside a polarized molecule (*24*–*26*, *27*, *29*). Molecular beam experiments have achieved high statistical sensitivity by measuring a large number of molecules over a ≈1-ms coherence time (*24*, *25*), whereas molecular ion–based experiments have used long Ramsey interrogation times (≈1 s), though with lower numbers (*26*, *27*, *29*). Measurements with trapped neutral polyatomic molecules can potentially combine the best features of each approach to achieve orders-of-magnitude improved statistical sensitivity (*6*).

[1]Department of Physics, Harvard University, Cambridge, MA 02138, USA. [2]Harvard-MIT Center for Ultracold Atoms, Cambridge, MA 02138, USA. [3]Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA.
*Corresponding author. Email: anderegg@g.harvard.edu (L.A.); hutzler@caltech.edu (N.R.H.)

Experimental approach for polyatomic molecule control

In this work, we demonstrate full quantum control over the internal states of a trapped polyatomic molecule in a vibrational bending mode with high polarizability in small electric fields. The protocol starts with preparing ultracold, optically trapped molecules in a single hyperfine level, after which a static electric field is applied to polarize the molecules. The strength of the polarizing electric field is tuned to obtain near-zero *g*-factor spin states, which have strongly suppressed sensitivity to magnetic field noise while retaining eEDM sensitivity. Microwave pulses are applied to create a coherent superposition of these zero *g*-factor spin states, which precesses under the influence of an external magnetic field. The precession phase is then read out by a combination of microwave pulses and optical cycling.

We observed spin precession over a range of electric and magnetic fields and characterized the present limitations to the coherence time of the measurement. With readily attainable experimental parameters, coherence times on the order of the state lifetime (>100 ms) could be realistically achieved. We therefore realized the key components of an eEDM measurement in this system. Note that eEDM sensitivity arises from the relativistic motion of the electron, which is enhanced by higher-mass nuclei. Although the light mass of CaOH precludes a competitive eEDM measurement (*30*), the protocol demonstrated here is directly transferable to heavier laser-cooled alkaline earth monohydroxides with identical internal level structures, such as SrOH, YbOH, and RaOH, which have substantially enhanced sensitivity to the eEDM (*6*, *11*, *12*, *30*, *31*).

In eEDM measurements with polarized molecules, the electron spin $\vec{S}$ precesses under the influence of an external magnetic field B_Z and the internal electric field of the molecule E_{eff}, which can be large owing to relativistic effects. Time evolution is described by the Hamiltonian

$$H = g_S \mu_B B_Z \vec{S} \cdot \hat{Z} - d_e E_{\mathrm{eff}} \vec{S} \cdot \hat{n}$$

$$= g_S \mu_B B_Z M_S - d_e E_{\mathrm{eff}} \Sigma \qquad (1)$$

Here, $g_S \approx 2$ is the electron spin *g*-factor, μ_B is the Bohr magneton, B_Z points along the lab $\hat{Z}$ axis, and the internal field E_{eff} points along the molecule's internuclear axis $\hat{n}$. We define the quantities $M_S = \vec{S} \cdot \hat{Z}$ and $\Sigma = \vec{S} \cdot \hat{n}$ to describe the electron's magnetic sensitivity and EDM sensitivity, respectively. The effect of the eEDM can be isolated by switching the orientation of the applied magnetic field or, alternatively, by switching internal states to change the sign of M_S or Σ. Performing both switches is a powerful technique for suppressing systematic errors (*25*, *26*).

Present EDM bounds rely on specific states in diatomic molecules that have an unusually small *g*-factor, which reduces sensitivity to stray magnetic fields (*24*, *26*). However, CaOH, like other laser-coolable molecules with structure amenable to eEDM searches (*6*, *31*–*33*), has a single valence electron, which results in large magnetic *g*-factors. In this work, we engineered reduced magnetic sensitivity by using an applied electric field E_Z to tune M_S to a zero-crossing while maintaining substantial eEDM sensitivity Σ. This technique is generic to polyatomic molecules with parity doublets. Details of a specific $M = \pm 1$ pair of zero *g*-factor states are shown in Fig. 1, A and B, with further information provided in (*34*). We emphasize that near the zero *g*-factor crossing, the eEDM sensitivity is nearly maximal because of the large projection of the electron spin onto the internal electric field of the molecule. Sensitivity to transverse magnetic fields is also suppressed in these zero *g*-factor states (*34*).

Experimental overview and single-state preparation

The experiment begins with laser-cooled CaOH molecules loaded from a magneto-optical trap (*14*) into an optical dipole trap (ODT) formed by a 1064-nm laser beam with a 25-μm waist

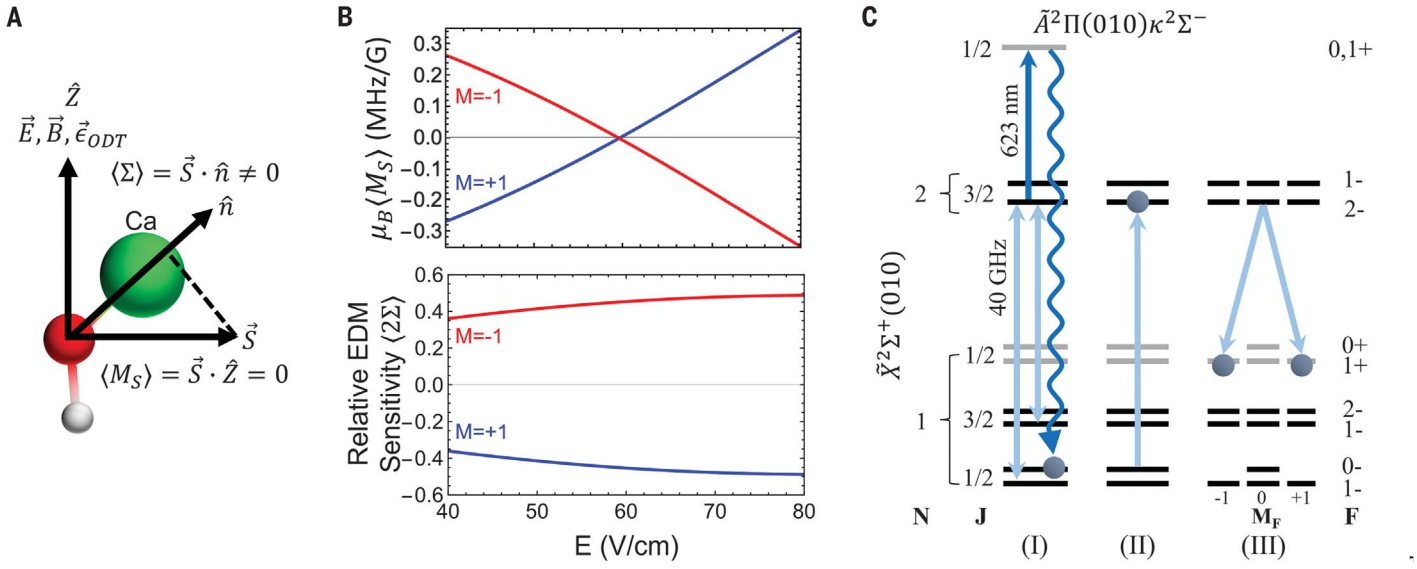


Fig. 1. Overview of the experiment. (**A**) A geometric picture of the bending molecule at the zero *g*-factor crossing, showing that the electron spin ($\vec{S}$) has a finite projection on the molecule axis ($\hat{n}$), giving eEDM sensitivity. However, the electron spin ($\vec{S}$) is orthogonal to the magnetic field ($\vec{B}$), which results in suppressed magnetic field sensitivity. (**B**) The magnetic sensitivity (top) and eEDM sensitivity (bottom) for a pair of zero *g*-factor states ($N = 1$, $J = 1/2^+$, $F = 1$, $M_F = \pm1$) are shown as a function of the applied electric field. (**C**) Experimental sequence to prepare the eEDM-sensitive state. First, the molecules are pumped into a single quantum state ($N = 1$, $J = 1/2^-$, $F = 0$) with a combination of microwave drives and optical pumping (I). Next, a microwave π-pulse drives the molecules into the $N = 2$, $J = 3/2^-$, $F = 2$, $M_F = 0$ state (II). Lastly, the eEDM measurement state is prepared as a coherent superposition of the $N = 1$, $J = 1/2^-$, $F = 1$, $M_F = \pm1$ states with a microwave π-pulse (III). The states that are optically detectable with the detection light are shown in black, whereas those not addressed by the detection light are in gray.

size, as described in previous work (*15*). The ODT is linearly polarized, and its polarization vector $\vec{\epsilon}_{ODT}$ defines the $\hat{Z}$ axis, along which we also apply magnetic and electric fields, $\vec{B} = B_Z\hat{Z}$ and $\vec{E} = E_Z\hat{Z}$, respectively, as depicted in Fig. 1A. We first nondestructively image the molecules in the ODT for 10 ms as normalization against variation in the number of trapped molecules. The molecules are then optically pumped into the $N = 1^-$ levels of the $\tilde{X}^2\Sigma^+(010)$ vibrational bending mode (*15*) (Fig. 1C), and the trap depth is adiabatically lowered by 3.5 times to reduce the effect of ac Stark shifts from the trap light and to lower the temperature of the molecules to 34 μK. Any molecules that were not pumped into $N = 1^-$ levels of the bending mode are heated out of the trap with a pulse of resonant laser light.

After transfer to the $\tilde{X}^2\Sigma^+(010)(N = 1^-)$ state, the molecular population is initially spread across 12 hyperfine Zeeman sublevels in the spin-rotation components $J = 1/2$ and $3/2$. To prepare the molecules in a single hyperfine state, we used a combination of optical pumping and microwave pulses, as shown in Fig. 1C. We first applied microwaves from the $N = 1$, $J = 3/2^-$ state up to the $N = 2$, $J = 3/2^-$ state. Because this transition is parity-forbidden, we applied a small electric field $E_Z = 7.5$ V/cm to slightly mix the parity of the $N = 1$ levels and provide transition strength. From the $N = 2$ state, we drove an optical transition to the excited $\tilde{A}^2\Pi(010)\kappa^2\Sigma^{(-)}, J = 1/2^+$ state. This state predominately decays to both the $F = 0$ (the target state) and $F = 1$ states in the $N = 1$, $J = 1/2^-$ manifold. After 3 ms of optical pumping, the microwaves were switched to drive the accumulated $N = 1$, $J = 1/2^-$, $F = 1$ population to the same $N = 2$, $J = 3/2^-$ state in $\tilde{X}(010)$, where they are excited by the optical light and pumped into the target $F = 0$ state. Once this optical pumping sequence is complete, we adiabatically ramped the electric field to $E_Z = 150$ V/cm to substantially mix parity, then drove the population up to the $N = 2$, $J = 3/2^-$, $F = 2$, $M = 0$ state with a microwave π-pulse (Fig. 1C, II). We cleaned out any remaining population in the $N = 1$ state with a depletion laser that resonantly drives the population to undetected rotational levels.

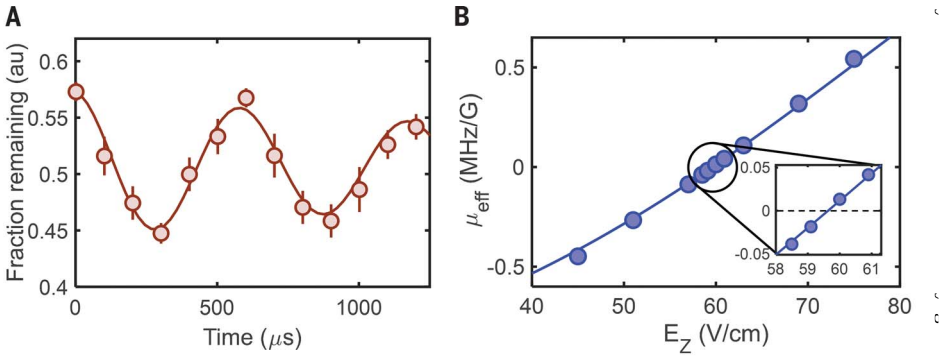


Fig. 2. Spin precession. (**A**) Spin precession of the eEDM-sensitive state in the presence of a bias magnetic field. Error bars represent 68% confidence intervals. au, arbitrary units. (**B**) Magnetic-field sensitivity of the eEDM state in CaOH as a function of electric field. The field sensitivity is determined by measuring the spin-precession frequency at different electric fields with an applied magnetic field of $B_Z = 110$ mG. Error bars are smaller than the markers. The solid curve is the calculated magnetic-field sensitivity in the presence of trap shifts using known molecular parameters (*34*).

Spin precession in an eEDM-sensitive state

To perform spin precession in the eEDM-sensitive state, we first adiabatically ramped the electric field to a value E_Z, then turned on a small bias magnetic field B_Z. We measured the electron spin precession frequency using a procedure analogous to Ramsey spectroscopy (*24*, *25*). The molecules were prepared by driving a π-pulse (2.5 μs), with microwaves linearly

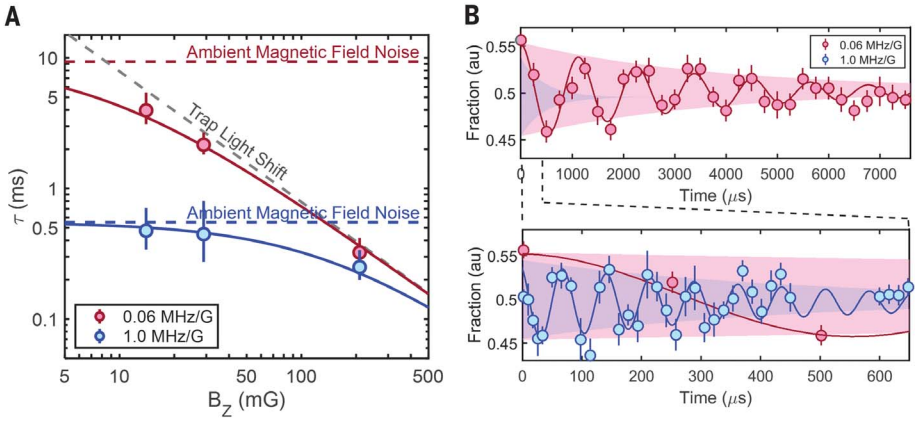


Fig. 3. Coherence time of the spin-precession signal. (**A**) Measured coherence times τ versus B_Z at different electric fields (red and blue markers, which correspond to different magnetic field sensitivities). The coherence time scales as $1/B_Z$ owing to ac Stark-shift broadening then plateaus at a limit set by the magnetic field instability δB. This limit increases as the *g*-factor approaches zero. Solid and dashed curves are fit to the data. The ambient magnetic field noise determined from the fit is $\delta B = 4^{+2}_{-1}$ mG, and the fitted decoherence time due to light shifts is $\tau = (1/B_Z) \times 80^{+20}_{-10}$ ms × mG. (**B**) Spin-precession signal, at B_Z = 15 mG, near the zero *g*-factor crossing (0.06 MHz/G; red) and far from the crossing (1.0 MHz/G; blue). Shaded regions indicate the fitted exponential decay envelope of the oscillations; the 1.0 MHz/G data are excluded from the top panel for clarity. The spin-precession coherence time is extended by 16 times by approaching the zero *g*-factor point.

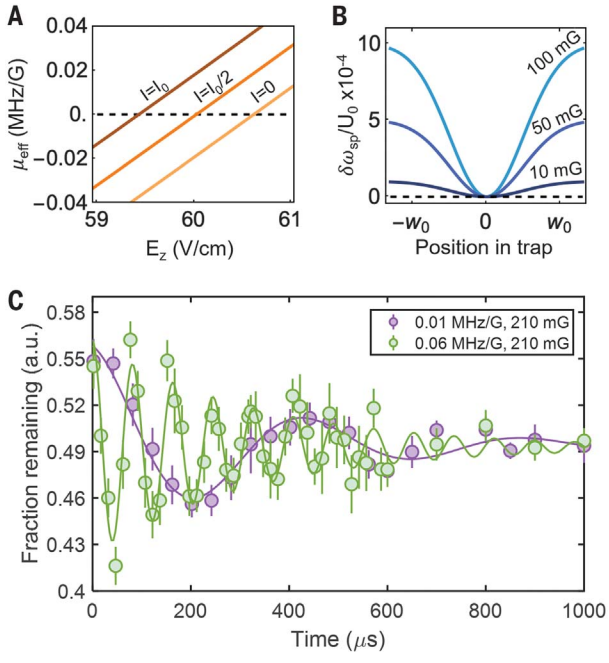


Fig. 4. Effect of trap light on coherence time. (**A**) Effective magnetic moment, μ_{eff}, as a function of electric field for several trap intensities *I*. The trap light shifts the location of the zero crossing in μ_{eff}. As a result, molecules at a finite temperature explore different magnetic-field sensitivities μ_{eff}. (**B**) Dependence of the spin-precession frequency (scaled by the trap depth U_0) on the position within the trap, where w_0 is the trap waist. At lower magnetic fields, the relative change in spin-precession frequency is reduced. (**C**) Two spin-precession curves taken at the same magnetic field (B_Z = 210 mG) but at different electric fields, showing that the ac Stark-shift limitation is independent of the effective *g*-factor because ac Stark shifts dominate the coherence time for large bias fields.

polarized along the lab $\hat{X}$ axis, into the "bright" superposition state $|B\rangle = (|M = 1\rangle + |M = -1\rangle)/\sqrt{2}$ within the N = 1, $J = 1/2^+$, F = 1, M = ±1 eEDM-sensitive manifold (Fig. 1C, III). The state begins to oscillate between the bright state and the "dark" state $|D\rangle = (|M = 1\rangle - |M = -1\rangle)/\sqrt{2}$ at a rate $\omega_{SP} = \mu_{eff} B_Z$, where the effective magnetic moment $\mu_{eff} = \mu_B g_{eff} = g_E \mu_B (\langle M_S \rangle_{M=1} - \langle M_S \rangle_{M=-1})$ is tuned by means of the applied electric field E_Z (Fig. 1B). The contribution from the $d_e E_{eff}$ term in Eq. 1 is negligible in CaOH but could be measured in heavier molecules with much larger E_{eff}. After a given time, a second π-pulse was applied to stop spin precession and transfer the bright state to the optically detectable $N = 2$, $J = 3/2^-$ level. Once the electric field was ramped down, the population remaining in the eEDM manifold, which has the opposite parity, is not optically detectable. We then imaged the ODT again and took the ratio of the first and second images (Fig. 2A). At long spin precession times (>10 ms), losses from background gas collisions (~1 s), blackbody excitation (~1 s), and the spontaneous lifetime of the bending mode (~0.7 s) lead to an overall loss of signal, as characterized in (*15*). This effect is mitigated with a fixed duration between the first and second images, making the loss independent of the precession time.

To map out the location of the zero *g*-factor crossing, we performed spin precession measurements at a fixed magnetic field B_Z = 110 mG for different electric fields (Fig. 2B). The spin precession frequency corresponds to an effective *g*-factor at that electric field. We found that the zero *g*-factor crossing within the N = 1, $J = 1/2^+$, F = 1, M = ±1 eEDM manifold occurs at an electric field of 59.6 V/cm, in agreement with theory calculations described in (*34*). We note that there is another zero *g*-factor crossing for the N = 1, $J = 3/2^+$, F = 1 manifold at ≈64 V/cm, which has a smaller eEDM sensitivity but the opposite slope of g_{eff} versus E_Z, thereby providing a powerful resource to reject systematic errors related to imperfect field reversals (*34*). We emphasize that although the location of these crossings is dependent on the structure of a specific molecule, their existence is generic in polyatomic molecules, which naturally have parity-doublet structure (*6*).

Coherence time and limitations

A critical component of the spin precession measurement is the coherence time, which sets the sensitivity of an eEDM search. Figure 3A shows the measured coherence time of our system at different applied fields B_Z and E_Z. We characterized two dominant limitations that wash out oscillations at long times. Variations in the spin precession frequency can be linearly expanded as $\delta\omega_{SP} = \mu_{eff}(\delta B_Z) + (\delta\mu_{eff})B_Z$. The first term describes magnetic field noise and drift of the applied bias field, given by δB_Z. The second term describes noise and drifts in the *g*-factor, δg_{eff}, which can arise from instability in the applied electric field, E_Z, or from ac Stark shifts (described below). Drifts in the bias electric field E_Z were found to be negligible in our apparatus.

Decoherence caused by magnetic field noise, δB_Z, is independent of the applied magnetic field but is proportional to μ_{eff} and can be mitigated by operating near the zero *g*-factor crossing. As shown in Fig. 3B, at an electric field of 90 V/cm, corresponding to a large magnetic moment of μ_{eff} = 1.0 MHz/G, we realized a magnetic field noise–limited coherence time of 0.5 ms at $B_Z \approx 15$ mG (blue points). At an electric field of 61.5 V/cm, corresponding to μ_{eff} = 0.06 MHz/G, which is much closer to the zero *g*-factor location, we found a coherence time of 4 ms at the same B_Z (red points in Fig. 3B).

At higher magnetic fields, the primary limitation to the coherence time is ac Stark shifts from the optical trapping light (Fig. 4). The intense *Z*-polarized ODT light leads to a shift

in the electric field at which the zero g-factor crossing occurs. Owing to the finite temperature of the molecules within the trap, they will explore different intensities of trap light and hence have different values of g_{eff}. The spread δg_{eff} causes variation of ω_{SP}, which leads to decoherence. In contrast to the magnetic field noise term, this effect is independent of the electric field E_Z but decreases monotonically with B_Z, which scales the frequency sensitivity to g-factor variations, $\delta\omega_{\mathrm{SP}} = B_Z\delta\mu_{\mathrm{eff}}$. The insensitivity of g-factor broadening to the exact value of g_{eff} is demonstrated in Fig. 4C. Decoherence caused by ac Stark shifts can be reduced by cooling the molecules to lower temperatures or by decreasing B_Z. The bias magnetic field can be reduced arbitrarily far until either transverse magnetic fields or magnetic field noise becomes dominant. From the decoherence rates measured in this work, it is expected that ac Stark shift–limited coherence times of ~1 s could be achieved at bias fields of $B_Z \sim 100$ μG.

From the above discussion, it is expected that the longest achievable coherence times will occur for very small g-factors, $g_{\mathrm{eff}} \approx 0$, and very small bias fields, $B_Z \approx 0$. Minimizing B_Z requires reducing the effects of both magnetic field noise and transverse magnetic fields to well below the level of the bias-field energy shifts. We canceled the transverse magnetic fields to below 1 mG by maximizing the spin precession period under the influence of transverse B fields only, and actively monitored and fed back on the magnetic field along each axis to minimize noise and drifts in B_Z. Note that the stainless-steel vacuum chamber has no magnetic shielding, which leads to high levels of magnetic field noise that would not be present in an apparatus designed for an eEDM search. Even under these conditions, we achieved a coherence time of 30 ms at an electric field of 60.3 V/cm (corresponding to $\mu_{\mathrm{eff}} = 0.02$ MHz/G) and a bias field of $B_Z \approx 2$ mG (*34*). However, at such a low bias field, the molecules are sensitive to 60-Hz magnetic field noise that is present in the unshielded apparatus, which is on the same order as the bias field. Because the experiment is phase-stable with respect to the ac line frequency, this 60-Hz magnetic field fluctuation causes a time-dependent spin precession frequency. Nevertheless, our prototype experiment confirms that long coherence times are possible. Any future eEDM experiment would have magnetic shielding that would greatly suppress nefarious magnetic fields from the environment. Such shielding could readily enable coherence times that exceed that of the ~0.5-s lifetime of the bending modes of similar linear polyatomic molecules with larger eEDM sensitivity (*15*).

Discussion and outlook

We have realized coherent control of optically trapped polyatomic molecules and demonstrated a realistic experimental roadmap for future eEDM measurements. By leveraging the distinctive features of the quantum levels in polyatomic molecules, we achieved a coherence time of 30 ms for paramagnetic molecules in a stainless-steel chamber with no magnetic shielding. With common shielding techniques used in past EDM experiments, there is a clear path to reducing stray fields and extending coherence times to >100 ms. At such a level, the dominant limitation becomes the finite lifetime of the bending mode (*15*). Even longer coherence times are possible with the right choice of parity-doublet states, as found in symmetric or asymmetric top molecules (*6*, *13*, *35*, *36*).

Following this roadmap with heavier trapped polyatomic molecules has the potential to provide orders-of-magnitude improvements to present bounds on T-violating physics. Using a recent study of the $\tilde{X}(010)$ state in YbOH (*37*), we identified similar $N = 1$ zero g-factor states for eEDM measurements with greatly improved sensitivity. In addition to the g-factor tuning demonstrated in this work, polyatomic molecules provide the ability to reverse the sign of Σ without reversing M_S, which is a crucial feature of recent experiments that has greatly improved the limit on the eEDM (*25*, *27*). For example, in the $N = 1$ manifold of CaOH, there is another zero g-factor crossing at a nearby electric field value, with 69% smaller values of Σ and opposite sign. Because the ratio of eEDM sensitivity to g-factor versus E_Z slope differs between these two crossings, measurements at both points could be used to suppress systematics caused by nonreversing fields that couple to the electric field dependence of the g-factor (*25*).

This work provides an experimental demonstration of the advantages of the rich level structure of polyatomic molecules for precision measurements. Although we focused here on spin precession with T-reversed states ($M = \pm 1$), many levels of interest can be favorably engineered for precision measurement experiments. In a recent proposal (*9*), parity doublets, magnetically tuned to degeneracy in optically trapped polyatomic molecules, were shown to be advantageous for searches for parity-violating physics. In another recent work (*7*), a microwave clock between rovibrational states in SrOH was proposed as a sensitive probe of ultralight dark matter, by using transitions tuned to electric and/or magnetic insensitivity. In these proposals, and as now experimentally demonstrated in our work, coherent control and state engineering in polyatomic molecules can mitigate systematic errors and enable robust searches for new physics.

REFERENCES AND NOTES

1. M. L. Wall, K. Maeda, L. D. Carr, *Ann. Phys.* **525**, 845–865 (2013).
2. M. Wall, K. Maeda, L. D. Carr, *New J. Phys.* **17**, 025001 (2015).
3. Q. Wei, S. Kais, B. Friedrich, D. Herschbach, *J. Chem. Phys.* **135**, 154102 (2011).
4. P. Yu, L. W. Cheuk, I. Kozyryev, J. M. Doyle, *New J. Phys.* **21**, 093049 (2019).
5. L. D. Augustovičová, J. L. Bohn, *New J. Phys.* **21**, 103022 (2019).
6. I. Kozyryev, N. R. Hutzler, *Phys. Rev. Lett.* **119**, 133002 (2017).
7. I. Kozyryev, Z. Lasner, J. M. Doyle, *Phys. Rev. A* **103**, 043313 (2021).
8. N. R. Hutzler, *Quantum Sci. Technol.* **5**, 044011 (2020).
9. E. B. Norrgard *et al*., *Commun. Phys.* **2**, 77 (2019).
10. Y. Hao *et al*., *Phys. Rev. A* **102**, 052828 (2020).
11. I. Kozyryev *et al*., *Phys. Rev. Lett.* **118**, 173201 (2017).
12. B. L. Augenbraun *et al*., *New J. Phys.* **22**, 022003 (2020).
13. D. Mitra *et al*., *Science* **369**, 1366–1369 (2020).
14. N. B. Vilas *et al*., *Nature* **606**, 70–74 (2022).
15. C. Hallas *et al*., *Phys. Rev. Lett.* **130**, 153202 (2023).
16. M. Zeppenfeld *et al*., *Nature* **491**, 570–573 (2012).
17. A. Prehn, M. Ibrügger, R. Glöckner, G. Rempe, M. Zeppenfeld, *Phys. Rev. Lett.* **116**, 063005 (2016).
18. D. DeMille, J. M. Doyle, A. O. Sushkov, *Science* **357**, 990–994 (2017).
19. T. E. Chupp, P. Fierlinger, M. J. Ramsey-Musolf, J. T. Singh, *Rev. Mod. Phys.* **91**, 015001 (2019).
20. M. S. Safronova *et al*., *Rev. Mod. Phys.* **90**, 025008 (2018).
21. C. Cesarotti, Q. Lu, Y. Nakai, A. Parikh, M. Reece, *J. High Energy Phys.* **2019**, 59 (2019).
22. M. Pospelov, A. Ritz, *Ann. Phys.* **318**, 119–169 (2005).
23. J. J. Hudson *et al*., *Nature* **473**, 493–496 (2011).
24. J. Baron *et al*., *Science* **343**, 269–272 (2014).
25. ACME Collaboration, *Nature* **562**, 355–360 (2018).
26. W. B. Cairncross *et al*., *Phys. Rev. Lett.* **119**, 153001 (2017).
27. T. S. Roussy *et al*., *Science* **381**, 46–50 (2023).
28. R. Alarcon *et al*., arXiv:2203.08103 [hep-ph] (2022).
29. Y. Zhou *et al*., *Phys. Rev. Lett.* **124**, 053201 (2020).
30. K. Gaul, R. Berger, *Phys. Rev. A* **101**, 012508 (2020).
31. T. A. Isaev, A. V. Zaitsevskii, E. Eliav, *J. Phys. At. Mol. Opt. Phys.* **50**, 225101 (2017).
32. T. A. Isaev, R. Berger, *Phys. Rev. Lett.* **116**, 063006 (2016).
33. I. Kozyryev, L. Baum, K. Matsuda, J. M. Doyle, *ChemPhysChem* **17**, 3641–3648 (2016).
34. See supplementary materials.
35. B. L. Augenbraun, J. M. Doyle, T. Zelevinsky, I. Kozyryev, *Phys. Rev. X* **10**, 031022 (2020).
36. B. L. Augenbraun *et al*., *Phys. Rev. A* **103**, 022814 (2021).
37. A. Jadbabaie *et al*., *New J. Phys.* **25**, 073014 (2022).
38. L. Anderegg *et al*., Data for quantum control of trapped polyatomic molecules for eEDM searches. Zenodo (2023); https://doi.org/10.5281/zenodo.8384389.

ACKNOWLEDGMENTS

A.J. acknowledges helpful discussions with C. Zhang and P. Yu. **Funding:** This work was supported by the Air Force Office of Scientific Research and the National Science Foundation (NSF). L.A. acknowledges support from the Harvard Quantum Initiative. N.B.V. from the Department of Defense National Defense Science and Engineering Graduate fellowship program, and P.R. from the NSF Graduate Research Fellowship Program. N.R.H. and A.J. acknowledge support from NSF CAREER program (PHY-1847550), the Gordon and Betty Moore Foundation (GBMF7947), and the Alfred P. Sloan Foundation (G-2019-12502). **Author contributions:** L.A., N.B.V., C.H., P.R., and A.J. performed the experiment and analyzed the data. N.R.H. and J.M.D. directed the study. All authors discussed the results and contributed to the manuscript. **Competing interests:** None declared. **Data and materials availability:** All data needed to evaluate the conclusions in this paper are present in the paper or in the supplementary materials. All data presented in this paper are deposited at Zenodo (*38*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adg8155
Supplementary Text
Figs. S1 to S4
References (*39*–*50*)

Submitted 23 January 2023; accepted 29 September 2023
10.1126/science.adg8155

MACHINE LEARNING

Prediction-powered inference

Anastasios N. Angelopoulos*†, Stephen Bates*†, Clara Fannjiang*†, Michael I. Jordan*†, Tijana Zrnic*†

Prediction-powered inference is a framework for performing valid statistical inference when an experimental dataset is supplemented with predictions from a machine-learning system. The framework yields simple algorithms for computing provably valid confidence intervals for quantities such as means, quantiles, and linear and logistic regression coefficients without making any assumptions about the machine-learning algorithm that supplies the predictions. Furthermore, more accurate predictions translate to smaller confidence intervals. Prediction-powered inference could enable researchers to draw valid and more data-efficient conclusions using machine learning. The benefits of prediction-powered inference were demonstrated with datasets from proteomics, astronomy, genomics, remote sensing, census analysis, and ecology.

Imagine a scientist has a machine-learning system that can supply accurate predictions about a phenomenon far more cheaply than any gold-standard experimental technique. The scientist may wish to use these predictions as evidence in drawing scientific conclusions. For example, accurate predictions of three-dimensional structures have been made for a vast catalog of known protein sequences (*1*, *2*) and are now being used in proteomics studies (*3*, *4*). Such machine-learning systems are increasingly common in modern scientific inquiry, in domains ranging from cancer prognosis to microclimate modeling. Predictions are not perfect, however, and this may lead to incorrect conclusions. Moreover, as predictions beget other predictions, the cumulative effect can amplify the imperfections. How can modern science leverage machine-learning predictions in a statistically principled way?

One way to use predictions is to follow the imputation approach: Proceed as if they are gold-standard measurements. Although this lets the scientist draw conclusions cheaply and quickly owing to the high-throughput nature of the machine-learning system, the conclusions may be invalid because the predictions may have biases.

Another possibility is to apply the classical approach: Ignore the machine-learning predictions and only use the available gold-standard measurements, which are typically far less abundant than predictions. The resulting discoveries will be statistically valid, but the smaller amount of data will limit the scope of possible discoveries.

This manuscript presents prediction-powered inference, a framework that achieves the best of both worlds: extracting information from the predictions of a high-throughput machine-learning system and guaranteeing statistical validity of the resulting conclusions. Prediction-powered inference provides a protocol for combining predictions, which are abundant but not always trustworthy, with gold-standard data, which are trusted but scarce, to compute confidence intervals and P values. The resulting confidence intervals and P values are statistically valid, as in the classical approach, but also leverage the information contained in the predictions, as in the imputation approach, to make the confidence intervals smaller and the P values more powerful.

Prediction-powered inference applies to any machine-learning system; as such, it absolves the need for case-by-case analyses dependent on the machine-learning algorithm on hand. The proposed protocol thereby could enable researchers to report on and assess the evidence for their conclusions in a fully standardized way.

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. Email: angelopoulos@berkeley.edu (A.N.A); stephenbates@berkeley.edu (S.B.); clarafy@berkeley.edu (C.F.); michael_jordan@berkeley.edu (M.I.J.); tijana.zrnic@berkeley.edu (T.Z.)
†These authors contributed equally to this work.

Protocol for prediction-powered inference

The protocol for prediction-powered inference proceeds as follows. The scientist wishes to construct a confidence interval for a quantity θ^*, such as the mean outcome or a regression coefficient quantifying the statistical association between the outcome and a feature. Toward this goal, they have access to a small gold-standard dataset of features paired with outcomes, $(X, Y) = ((X_1, Y_1), ..., (X_n, Y_n))$, as well as the features of a large unlabeled dataset, $(X', Y') = \left((X'_1, Y'_1), ..., (X'_N, Y'_N)\right)$, where the true outcomes $Y'_1, ..., Y'_N$ are not observed. Typically, N is much larger than n . Both datasets are sampled at random from a larger population. Further, for both datasets the scientist has predictions of the outcomes made by a machine-learning algorithm based on the features, denoted $\left(\hat{Y}_1, ..., \hat{Y}_n\right)$ and $\left(\hat{Y}'_1, ..., \hat{Y}'_N\right)$, respectively. The following exposition focuses on confidence intervals; however, by the standard duality between confidence intervals and P values, the presented tools immediately carry over to valid P-value constructions and hypothesis tests; see supplementary materials (SM) for details.

Prediction-powered inference uses the g… standard dataset to quantify and correct… the errors made by the machine-learning algorithm on the unlabeled dataset, thereby enabling researchers to reliably incorporate predictions when constructing confidence intervals. The three-step protocol is outlined below and visualized in Fig. 1.

1) Estimand. The first step is to select an estimand θ^*. The estimand is the quantity the scientist is interested in knowing—for example, the mean outcome $E[Y_i]$, median outcome $\text{median}(Y_i)$, a linear regression coefficient obtained by regressing Y onto X, etc.

2) Measure of fit and rectifier. The key step is to identify the right measure of fit m_θ and rectifier Δ_θ for the selected estimand. For every candidate value of the estimand θ, the measure of fit m_θ is computed on the unlabeled dataset imputed with predictions, $(X', \hat{Y}')$ and quantifies how likely θ^* is to be equal to θ on the basis of the imputed data. The closer m_θ is to zero, the more plausible it is for θ^* to be equal to θ.

The rectifier Δ_θ is a notion of prediction error that is relevant for the estimand of interest. It is defined as the difference of the measure of fit m_θ computed on the labeled data, (X, Y), and the labeled data when the true outcomes are replaced with predicted ones, $\left(X, \hat{Y}\right)$. If the predictions are perfect, the rectifier is equal to zero.

Table 1 states the appropriate measure of fit and rectifier for common estimands of interest: the mean outcome, median outcome, q-quantile of the outcome, and linear and logistic regression coefficients when regressing Y onto X. A general recipe for deriving the right measure of fit and corresponding rectifier for a broad class of other estimands is provided in the SM.

3) Prediction-powered confidence interval. Finally, the measure of fit and rectifier are carefully combined to form a prediction-powered confidence interval for θ^*. This process is called rectifying the confidence interval. The prediction-powered confidence interval is constructed as $C^{\text{PP}} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$ and is guaranteed to contain the estimand with probability at least $1 - \alpha$. Here, $w_\theta(\alpha)$ is a constant that depends on the confidence level; it is explicitly stated in Theorem S1 in the SM.

Properties of prediction-powered inference

We proved mathematically that prediction-powered inference yields a confidence interval that contains the true value of the estimand at the desired confidence level, such as 95%. Notably, this validity is guaranteed for any machine-learning algorithm and any underlying data distribution. Similarly, the corresponding P values are also valid for any machine-learning algorithm and data distribution. See SM for the details of the mathematical proof

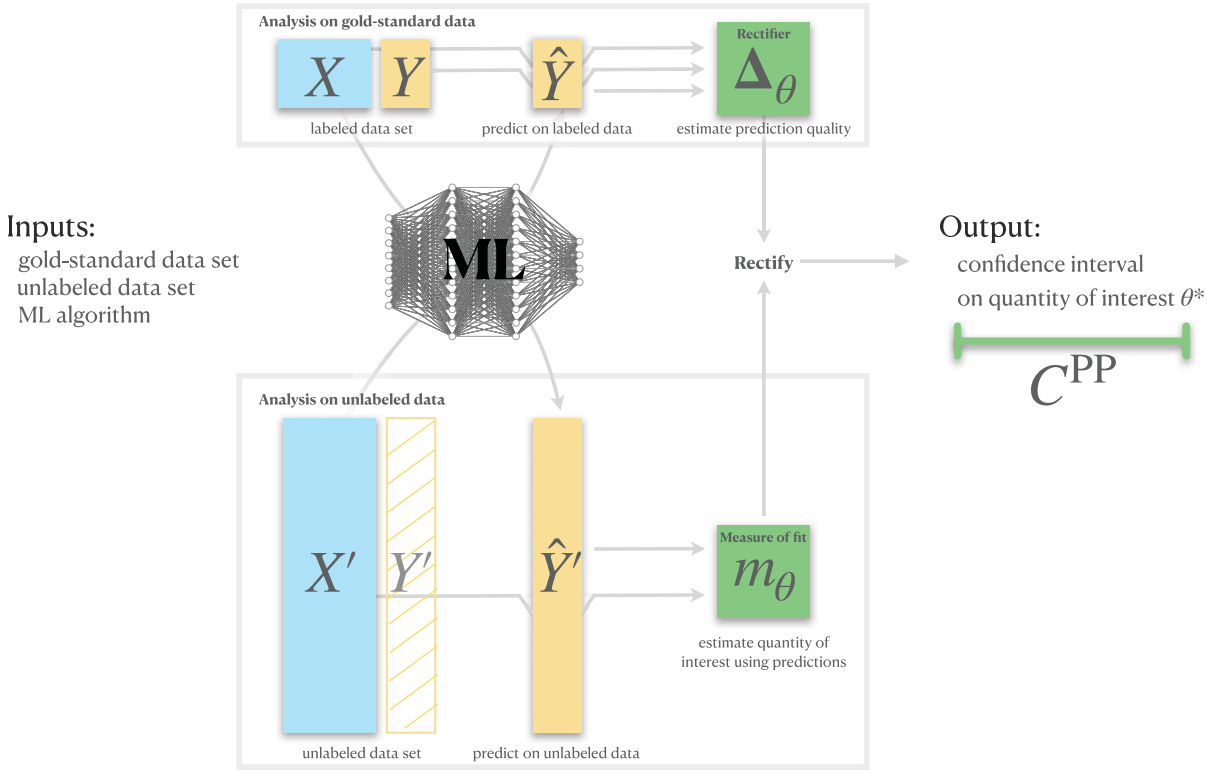


Fig. 1. Protocol for prediction-powered inference. The protocol is illustrated graphically as a block diagram. The inputs are the gold-standard dataset, the unlabeled dataset, and the machine-learning (ML) algorithm. The top block contains an analysis on gold-standard data, in which the rectifier, a measure of the prediction errors, is estimated using the labeled dataset. The bottom block contains an analysis on unlabeled data, wherein the quantity of interest is estimated using predictions. These analyses combine to form the prediction-powered confidence interval. For concrete examples of the rectifier and measure of fit, see Table 1. For a detailed theoretical exposition and more general definitions of these quantities, see SM.

of validity. A researcher relying on a deep neural network for predictions can therefore draw reliable conclusions, even though its predictions will inevitably be imperfect. Furthermore, prediction-powered inference enables more informative inferences than the classical approach, in which the researcher does not use machine-learning predictions: The confidence intervals are narrower, and the P values are more powerful. This is intuitive; prediction-powered inference carefully extracts information from the imputed data and thus has access to a larger sample size.

General applicability

Beyond quantities such as means, quantiles, and regression coefficients, the principle of prediction-powered inference can be used for constructing valid confidence intervals for any estimand that can be expressed as the minimizer of a convex objective function. This master protocol, which generalizes all the special cases instantiated in Table 1, is the core technical contribution of this work. We explained prediction-powered inference in greater generality and proved its validity in this general case in the SM. Because many important quantities can be expressed in terms of a convex-optimization problem, prediction-powered inference thus addresses many data-analysis goals beyond those explicitly demonstrated in this article.

Inference under distribution shift

Prediction-powered inference is also applicable to settings with distribution shift, i.e., the more challenging case where the unlabeled data are collected under different conditions than the gold-standard data. Two types of distribution shift are considered: label shift and covariate shift. The protocol retains the same properties as before: It is statistically valid for any machine-learning algorithm and boosts statistical power by making use of machine-learning predictions.

For covariate shift—the setting where only the feature distribution changes between the labeled and the unlabeled data—prediction-powered inference handles all estimation problems handled by the master protocol. This is done by appropriately reweighting the data; see Corollary S13 in the SM for details.

For label shift—the setting where only the label proportions change between the labeled and the unlabeled data—prediction-powered inference can be applied to estimands of the form $\theta^* = \mathrm{E}\left[\nu\left(Y_i'\right)\right]$, for a fixed function ν. For example, choosing $\nu(y) = 1\{y = k\}$ asks for inference on the proportion of instances that belong to class k. See Theorem S3 in the SM for a full description of the method.

Application of prediction-powered inference to real datasets

We demonstrated prediction-powered inference on several real tasks. In each, we computed a prediction-powered confidence interval for an estimand and compared it to intervals obtained through the classical approach and the imputation approach. In all cases, the imputation approach, which uses machine-learning predictions without accounting for prediction errors, did not contain the true value of the estimand. The widths of the two valid approaches, prediction-powered and classical, were compared as a function of the amount of labeled data used. In addition, we compared the number of labeled examples needed to reject a null hypothesis at level $1 - \alpha = 95\%$ with high probability. See (*5*) for a Python package implementing prediction-powered inference, which contains code for reproducing the experiments, and (*6*) for the data used in the experiments.

Relating protein structure and posttranslational modifications

The goal was to characterize whether various types of posttranslational modifications

Table 1. Prediction-powered inference for common statistical problems. Given a measure of fit m_θ and rectifier Δ_θ, prediction-powered inference computes a confidence interval as $C^{PP} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \le w_\theta(\alpha)\}$, where $w_\theta(\alpha)$ is a constant that depends on the error level α (see Theorem S1 in the SM). Algorithms S1 to S6 are stated in the SM. The last row ("convex minimizer") refers to a method that generalizes the methods in previous rows.

Estimand	Measure of fit m_θ	Rectifier Δ_θ	Procedure
Mean outcome	$\theta - \frac{1}{N}\sum_{i=1}^{N} \hat{Y}'_i$	$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)$	Alg. S1
Median outcome	$\frac{1}{2N}\sum_{i=1}^{N}\text{sign}\left(\theta - \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \le \theta\} - 1\left\{\hat{Y}_i \le \theta\right\}\right)$	Alg. S2
q-quantile of outcome	$-q + \frac{1}{N}\sum_{i=1}^{N} 1\left\{\hat{Y}'_i \le \theta\right\}$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \le \theta\} - 1\left\{\hat{Y}_i \le \theta\right\}\right)$	Alg. S3
Linear regression	$\theta - (X')^{+}\hat{Y}'$	$X^{+}(\hat{Y} - Y)$	Alg. S4
Logistic regression	$\frac{1}{N}\sum_{i=1}^{N} X'_i\left(\frac{1}{1+e^{-\theta^T X'_i}} - \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n} X_i\left(\hat{Y}_i - Y_i\right)$	Alg. S5
Convex minimizer	$\frac{1}{N}\sum_{i=1}^{N}\nabla L_\theta\left(X'_i, \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(\nabla L_\theta\left(X_i, \hat{Y}_i\right) - \nabla L_\theta(X_i, Y_i)\right)$	Alg. S6

(PTMs) occurred more frequently in intrinsically disordered regions (IDRs) of proteins (*7*). Recently, Bludau *et al.* (*3*) studied this relationship on an unprecedented proteome-wide scale by using structures predicted by AlphaFold (*1*) to predict IDRs, in contrast to previous work, which was limited to far fewer experimentally derived structures.

To quantify the association between PTMs and IDRs, the authors applied the imputation approach: They computed the odds ratio between AlphaFold-based IDR predictions and PTMs on a dataset of hundreds of thousands of protein sequence residues (*8*). Using prediction-powered inference, we could combine AlphaFold-based predictions together with gold-standard IDR labels to give a confidence interval for the true odds ratio that is statistically valid, in contrast with the interval constructed with the imputation approach, and smaller than the interval constructed using the classical approach. We used the fact that the odds ratio could be written in terms of two means and applied the recipe from the first row of Table 1; see SM for details.

We had 10,803 data points from Bludau *et al.* (*3*). For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and treated the remaining $N = 10,803 - n$ points as the unlabeled dataset for which we did not observe the IDR labels. For all values of n and all three different types of PTMs that we examined, the prediction-powered confidence intervals were smaller than classical intervals; see row A in Fig. 2. Often, the classical intervals were large enough that they contained the odds ratio value of one, which means the direction of the association could not be determined from the confidence interval. However, the imputed confidence interval was far too small and significantly overestimated the true odds ratio. To reject the null hypothesis that the odds ratio is no greater than one, prediction-powered inference required $n = 316$ labeled observations, and the classical approach required $n = 799$ labeled observations; see row A in Table 2.

Galaxy classification

The goal was to determine the demographics of galaxies with spiral arms, which are correlated with star formation in the disks of low-redshift galaxies, and therefore, contribute to the understanding of star formation in the Local Universe. A large citizen science initiative called Galaxy Zoo 2 (*9*) has collected human annotations of roughly 300,000 images of galaxies from the Sloan Digital Sky Survey (*10*) with the goal of measuring these demographics. We sought to explore the use of machine learning to improve the effective sample size and decrease the requisite number of human-annotated galaxies.

We focused on estimating the fraction of galaxies with spiral arms. We had 1,364,122 labeled galaxy images from Galaxy Zoo 2, from which we simulated labeled and unlabeled datasets as follows. For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and used the remaining $N = 1,364,122 - n$ points as the unlabeled dataset. We then used the first row of Table 1 to construct prediction-powered intervals. The prediction-powered confidence intervals for the mean were consistently much smaller than the classical intervals and they retained validity, and the imputation strategy failed to cover the ground truth; see Fig. 2, row B. To reject the null hypothesis that the fraction of galaxies with spiral arms is at most 0.2, prediction-powered inference required $n = 189$ labeled examples, and classical inference required $n = 449$ examples; see Table 2, row B.

Distribution of gene expression levels

Next, we constructed prediction-powered confidence intervals on quantiles that characterize how a population of promoter sequences affects gene expression. Recently, Vaishnav *et al.* (*11*) trained a state-of-the-art transformer model to predict the expression level of a particular gene induced by a promoter sequence. They used the model's predictions to study the effects of promoters—for example, by assessing how quantiles of predicted expression levels differ between different populations of promoters.

Here we focused on estimating different quantiles of gene expression levels induced by native yeast promoters. We had 61,150 labeled native yeast promoter sequences from Vaishnav *et al.* (*11*), from which we simulated labeled and unlabeled datasets as follows. For each of 100 trials, we randomly sampled n points to serve as the labeled dataset and used the remaining $N = 61,150 - n$ points as the unlabeled dataset. We then used the second and third row of Table 1 to construct prediction-powered intervals for the median, as well as the 25% and 75% quantiles, of the expression levels. The prediction-powered confidence intervals for all three quantiles were much smaller than the classical intervals for all values of n. See row C in Fig. 2 for the results for the median and fig. S6 for the other two quantiles. We also evaluated the number of labeled examples required by prediction-powered inference and classical inference, respectively, to reject the null hypothesis that the median gene expression level is at most five. Prediction-powered inference required $n = 764$ examples and classical inference required $n = 900$ examples; see row C in Table 2.

Estimating deforestation in the Amazon

The goal was to estimate the fraction of the Amazon rainforest lost between 2000 and 2015. Gold-standard deforestation labels for parcels of land are scarce, having been collected in large part through field visits, an expensive process not suited for large areas (*12*). However, machine-learning predictions of forest cover based on satellite imagery are readily available for the entire Amazon (*13*). We began with 1596 gold-standard deforestation labels for parcels of land in the Amazon. For each of 100 trials, we randomly sampled n data points to serve as the labeled dataset and used the remaining data points as the unlabeled dataset. We used the first row of Table 1 to construct the prediction-powered intervals. The imputation approach yielded a small confidence interval that failed to cover the true deforestation fraction. The classical

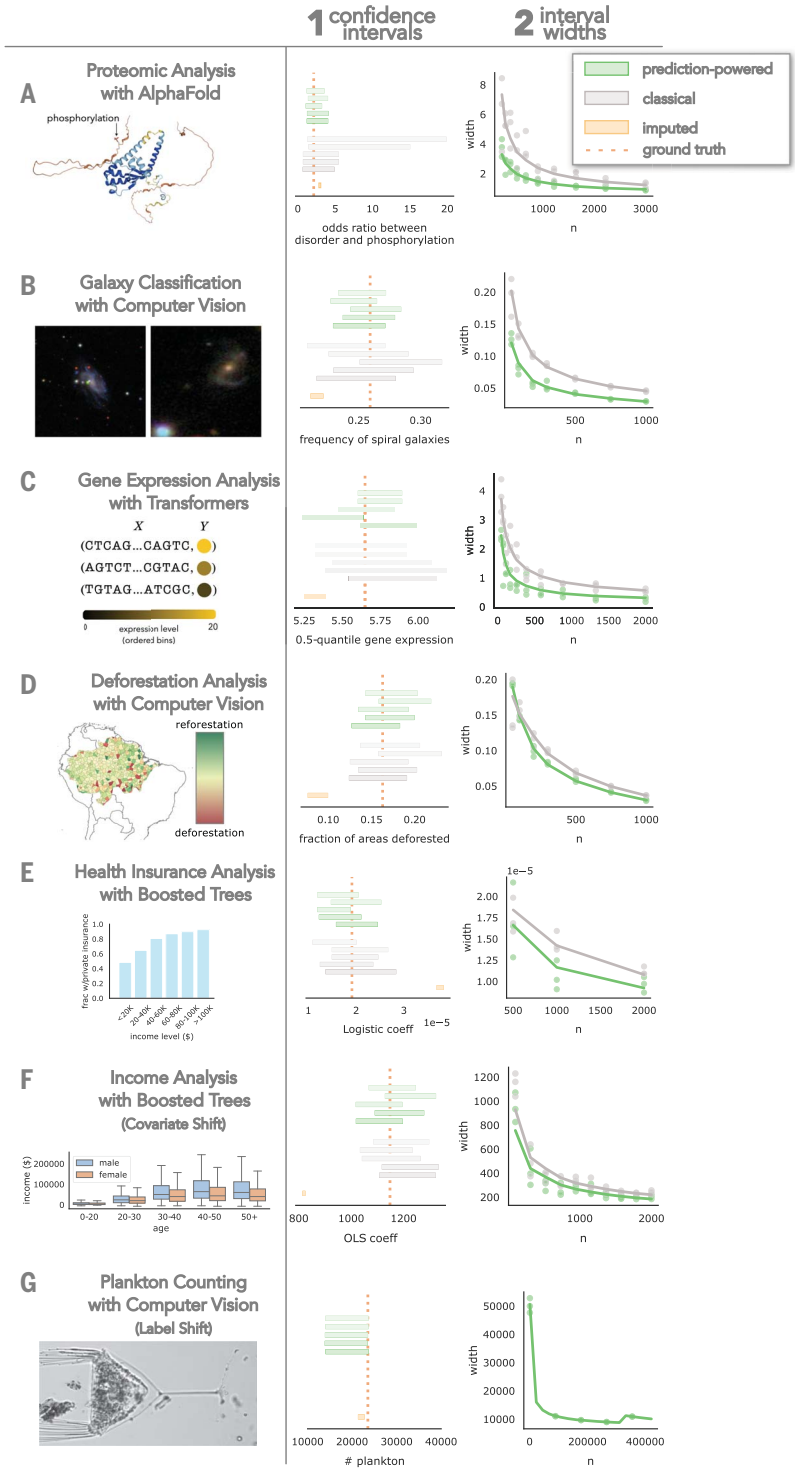


Fig. 2. Comparison of prediction-powered inference to the classical and imputation approaches on real tasks. Each row (**A** to **G**) is a different application domain. Panel 1 plots confidence intervals computed using the three approaches; for prediction-powered inference and the classical approach, intervals for five randomly chosen splits into labeled and unlabeled data are plotted. The value denoted as "ground truth" is the estimate computed on all $n + N$ data points (the true labels were available for all data points for the purpose of conducting the experiments). Panel 2 plots the average confidence interval width, as well as the width in five randomly chosen trials, for varying n, for prediction-powered inference and the classical approach; both are statistically valid solutions. The last problem setting (G) does not have a classical counterpart because the data are collected under distribution shift, hence the classical approach is not valid.

approach did cover the truth at the expense of a wider interval and, accordingly, diminished inferential power. The prediction-powered intervals were smaller than the classical intervals and retained validity; see row D in Fig. 2. We also compared the number of gold-standard deforestation labels required by prediction-powered inference and the classical approach to reject the null hypothesis that there is no deforestation. We obtain $n = 21$ labels for prediction-powered inference and $n = 35$ labels for the classical approach; see row D in Table 2.

Relationship between income and private health insurance

The goal was to investigate the quantitative effect of income on the procurement of private health insurance using US census data. Concretely, we used the Folktables interface (*14*) to download census data from California in the year 2019 (378,817 individuals). As the labeled dataset with the health insurance indicator, n census entries were randomly sampled. The remaining data were used as the unlabeled dataset. We used a gradient-boosted tree (*15*) trained on the previous year's data to predict the health insurance indicator in 2019. We constructed a prediction-powered confidence interval on the logistic regression coefficient using the fifth row of Table 1. Results in row E in Fig. 2 show that prediction-powered inference covered the ground truth, the classical interval was wider, and the imputation strategy failed to cover the ground truth. We also compared the number of gold-standard labels required by prediction-powered inference and the classical approach to reject the null hypothesis that the logistic regression coefficient is no greater than 1.5×10^{-5}. We observed a significant sample size reduction with prediction-powered inference, which required $n = 5569$ labels, whereas classical inference required $n = 6653$ labels.

Relationship between age and income in a covariate-shifted population

The goal was to investigate the relationship between age and income using US census data. The same dataset was used as in the previous experiment, but the features were age and sex, and the target was yearly income in dollars. Furthermore, a shift in the distribution of the covariates was introduced between the gold-standard and unlabeled datasets by randomly sampling the unlabeled dataset with sampling weights of 0.8 for females and 0.2 for males. We used a gradient-boosted tree (*15*) trained on the previous year's raw data to predict the income in 2019. We constructed a prediction-powered confidence interval on the ordinary least squares (OLS) regression coefficient using a covariate-shift robust version of prediction-powered inference, stated in Corollary S13 in the SM. Results in row F in Fig. 2 show that

Table 2. Number of labeled examples needed to make a discovery with prediction-powered inference and classical inference. The rows (A to F) correspond to the application domains from Fig. 2. For each application, a null hypothesis about θ^* is tested at level 95%. For details, see the SM.

Problem	Prediction-powered inference	Classical inference
A Proteomic analysis with AlphaFold	$n = 316$	$n = 799$
B Galaxy classification with computer vision	$n = 189$	$n = 449$
C Gene expression analysis with transformers	$n = 764$	$n = 900$
D Deforestation analysis with computer vision	$n = 21$	$n = 35$
E Health insurance analysis with boosted trees	$n = 5569$	$n = 6653$
F Income analysis with boosted trees	$n = 177$	$n = 282$

prediction-powered inference covered the ground truth, the classical interval was wider, and the imputation strategy failed to cover the ground truth. We also compared the number of gold-standard labels required by prediction-powered inference and the classical approach to reject the null hypothesis that the OLS regression coefficient is no greater than 800 . We observed a significant sample size reduction with prediction-powered inference, which required $n = 177$labels, whereas classical inference required $n = 282$ labels.

Counting plankton

Assessment of the increases in phytoplankton growth during springtime warming is important for the study of global biogeochemical cycling in response to climate change. We counted the number of plankton observed by the Imaging FlowCytobot (*16*, *17*), an automated, submersible flow cytometry system, at Woods Hole Oceanographic Institution in the year 2014. We had access to data from 2013, which were labeled, and we imputed the 2014 data with machine-learning predictions from a state-of-the-art ResNet fine-tuned on all data up to and including 2012. The features, X_i, are images of organic matter taken by the FlowCytobot and the labels, Y_i, are one of {detritus, plankton}, where detritus represents unspecified organic matter.

The labeled dataset consisted of 421,238 image–label pairs from 2013, and we received 329,832 labeled images from 2014. We used the data from 2014 as our unlabeled data and confirmed our results against those that were hand-labeled. The years 2013 and 2014 had a distribution shift, primarily caused by the change in the base frequency of plankton observations with respect to detritus. To apply prediction-powered inference to count the number of plankton recorded in 2014, we used the label-shift-robust technique described in Theorem S3 in the SM. The results in row G in Fig. 2 show that prediction-powered inference covered the ground truth and the imputation strategy failed to cover the ground truth.

Related work

Thematically, prediction-powered inference is most similar to the work of Wang *et al.* (*18*), who introduced a method to correct machine-learning predictions for the purpose of subsequent inference. However, this procedure is not guaranteed to provide coverage in general and requires strong assumptions about the relationship between the prediction model and the true response, whereas prediction-powered inference provides provably valid conclusions under minimal assumptions about the data-generating distribution.

There has been an increasing body of work on estimation with many unlabeled data points and few labeled data points (*19*–*27*), focusing on efficiency in semiparametric or high-dimensional regimes. Prediction-powered inference continues in this vein but focuses on the setting where the scientist has access to a good predictive model fit on separate data. This allows tackling a much wider range of estimands (e.g., minimizers of any convex objective) and gives valid inferences without assumptions about the machine-learning model. Second, prediction-powered inference goes beyond random sampling and applies to certain forms of distribution shift.

Prediction-powered inference is conceptually related to conformal prediction (*28*). Both methodologies leverage a predictive model and a labeled dataset. From this point on, however, the two methods diverge: Prediction-powered inference has additional unlabeled data and gives a confidence set that contains a population-level quantity such as the mean outcome with high probability; conformal prediction gives a confidence set for a test instance that contains the true label with high probability. Thus, the goals of prediction-powered inference and conformal prediction differ greatly from the statistical perspective. Furthermore, the mathematical tools used in the frameworks are entirely different, and neither method can be applied nontrivially to solve the objective of the other.

See SM for a further discussion of related work and the relationship of prediction-powered inference to existing baselines, as well as for empirical comparisons.

Conclusions

The past decade has witnessed rapid development and deployment of large-scale machine-learning systems across science. This surge is proceeding, however, with little statistical justification to allow these black-box systems to be used to draw scientific conclusions responsibly. Prediction-powered inference is a standardized protocol for constructing provably valid confidence intervals and P values, allowing the scientist to use the power and scale of machine-learning systems. On an array of scientific problems, we demonstrated that prediction-powered inference achieved high statistical power owing to the use of machine-learning predictions and retained statistical validity.

One question that remains open is how to handle more general forms of distribution shift. In practice, distribution shifts are often a result of a joint influence of several different forms of shift, including covariate shift and label shift and possibly others. Understanding how to handle such settings remains an important avenue for future work.

A limitation of prediction-powered inference is that it does not improve upon the classical approach when the predictions are not accurate enough or when the unlabeled dataset is not large enough compared to the gold-standard dataset. These points are demonstrated, both theoretically and empirically, in SM section "Cases Where Prediction-Powered Inference Is Underpowered." Nevertheless, given the growing number of settings with excellent predictive models and abundant unlabeled data, there is increasing potential for prediction-powered inference to benefit scientific research.

REFERENCES AND NOTES

1. J. Jumper *et al.*, *Nature* **596**, 583–589 (2021).
2. K. Tunyasuvunakool *et al.*, *Nature* **596**, 590–596 (2021).
3. I. Bludau *et al.*, *PLOS Biol.* **20**, e3001636 (2022).
4. I. Barrio-Hernandez *et al.*, Clustering predicted structures at the scale of the known protein universe. bioRxiv **2023-03** (2023).
5. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, ppi-py: A Python package for scientific discovery using machine learning, Zenodo (2023); https://doi.org/10.5281/zenodo.8403931.
6. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-Powered Inference: Data Sets, Zenodo (2023); https://doi.org/10.5281/zenodo.8397451.
7. L. M. Iakoucheva *et al.*, *Nucleic Acids Res.* **32**, 1037–1049 (2004).
8. UniProt Consortium, *Nucleic Acids Res.* **43**, D204–D212 (2015).
9. K. W. Willett *et al.*, *Mon. Not. R. Astron. Soc.* **435**, 2835–2860 (2013).
10. D. G. York *et al.*, *Astron. J.* **120**, 1579–1587 (2000).
11. E. D. Vaishnav *et al.*, *Nature* **603**, 455–463 (2022).
12. E. L. Bullock, C. E. Woodcock, C. Souza Jr., P. Olofsson, *Glob. Chang. Biol.* **26**, 2956–2969 (2020).
13. J. O. Sexton *et al.*, *Int. J. Digit. Earth* **6**, 427–448 (2013).
14. F. Ding, M. Hardt, J. Miller, L. Schmidt, in *Advances in Neural Information Processing Systems* **34** (2021), pp. 6478–6490.
15. T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

16. R. J. Olson, A. Shalapyonok, H. M. Sosik, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **50**, 301–315 (2003).
17. E. C. Orenstein, O. Beijbom, E. E. Peacock, H. M. Sosik, WHOI-Plankton- A large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv:1510.00745 [cs.CV] (2015).
18. S. Wang, T. H. McCormick, J. T. Leek, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30266–30275 (2020).
19. M. S. Pepe, *Biometrika* **79**, 355–365 (1992).
20. J. Lafferty, L. Wasserman, in *Advances in Neural Information Processing Systems* **20** (2007), pp. 801–808.
21. A. Zhang, L. D. Brown, T. T. Cai, *Ann. Stat.* **47**, 2538–2566 (2019).
22. A. Chakrabortty, G. Dai, E. Tchetgen Tchetgen, A general framework for treatment effect estimation in semi-supervised and high dimensional settings. arXiv:2201.00468 [stat.ME] (2022).
23. A. Chakrabortty, T. Cai, *Ann. Stat.* **46**, 1541–1572 (2018).
24. Y. Zhang, J. Bradic, *Biometrika* **109**, 387–403 (2022).
25. S. Deng, Y. Ning, J. Zhao, H. Zhang, Optimal and safe estimation for high-dimensional semi-supervised learning. arXiv:2011.14185 [stat.ME] (2020).
26. D. Azriel *et al.*, *J. Am. Stat. Assoc.* **117**, 2238–2251 (2022).
27. A. Chakrabortty, G. Dai, R. J. Carroll, Semi-supervised quantile estimation: robust and efficient inference in high dimensional settings. arXiv:2201.10208 [stat.ME] (2022).
28. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World* (Springer, 2005), vol. 5.

ACKNOWLEDGMENTS

We thank A. Kohli for suggesting the term "rectifier," E. Orenstein for helpful discussions related to the WHO-I Plankton dataset, and S. Wang for help with the remote-sensing experiment. **Funding:** Office of Naval Research grant N00014-21-1-2840 (M.I.J.); National Science Foundation Graduate Research Fellowship (A.N.A., C.F.).
Author contributions: Conceptualization: A.N.A., S.B., C.F., M.I.J., T.Z.; Methodology: A.N.A., S.B., C.F., M.I.J., T.Z.; Formal analysis: A.N.A., S.B., C.F., M.I.J., T.Z.; Funding acquisition: M.I.J.; Supervision: M.I.J.; Writing – original draft: A.N.A., S.B., C.F., M.I.J., T.Z.; Writing – review and editing: A.N.A., S.B., C.F., M.I.J., T.Z.
Competing interests: None declared. **Data and materials availability:** The data and code are available in the accompanying Python package (*5*) and data repository (*6*). All other data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIAL

science.org/doi/10.1126/science.adi6000
Supplementary Text
Figs. S1 to S6
References (*29*–*56*)

Submitted 7 May 2023; accepted 5 October 2023
10.1126/science.adi6000

PHAGE ECOLOGY

Rapid bacteria-phage coevolution drives the emergence of multiscale networks

Joshua M. Borin[1], Justin J. Lee[1], Adriana Lucia-Sanz[2], Krista R. Gerbino[1], Joshua S. Weitz[3,4,5]†, Justin R. Meyer[1]*

Interactions between species catalyze the evolution of multiscale ecological networks, including both nested and modular elements that regulate the function of diverse communities. One common assumption is that such complex pattern formation requires spatial isolation or long evolutionary timescales. We show that multiscale network structure can evolve rapidly under simple ecological conditions without spatial structure. In just 21 days of laboratory coevolution, *Escherichia coli* and bacteriophage Φ21 coevolve and diversify to form elaborate cross-infection networks. By measuring ~10,000 phage-bacteria infections and testing the genetic basis of interactions, we identify the mechanisms that create each component of the multiscale pattern. Our results demonstrate how multiscale networks evolve in parasite-host systems, illustrating Darwin's idea that simple adaptive processes can generate entangled banks of ecological interactions.

The evolution of life has yielded complex ecological networks encompassing everything from food webs in the savannah to pollinator networks in the rainforest and microbe interactions in the oceans. Notably, ecological networks in these vastly different environmental contexts show similar, repeatable patterns (*1*–*6*). Two commonly observed patterns include nestedness, where interactions between narrow-range specialists are hierarchically embedded or "nested" within the broader interaction ranges of generalists (*1*, *7*), and modularity in which interactions between species form within—but not between—groups, generating distinct "modules" (*2*, *6*). When large-scale analyses of natural community interactions are conducted multiscale networks are often uncovered, whereby modularity is observed at broad scales and nestedness is observed within modules (*3*–*5*, *8*, *9*). How these multiscale patterns emerge within ecological networks is an open question and the answer has implications for how complexity evolves in biological systems. It has been suggested that the formation of such intricate structures requires geographic separation (*8*, *10*, *11*), long evolutionary timescales (*12*, *13*), or other complex processes (*5*, *14*). However, given the repeatability of these structures in disparate systems, it has been hypothesized that endogenous processes within biological systems, such as coevolution between interacting species (including bacteria and their viral parasites), could generate such complexity even on short timescales and in closed ecological communities (*15*).

Bacteria and their viruses (phages) provide tractable systems for studying how ecological networks evolve (*4*, *16*, *17*). By isolating phages and bacteria from a given ecosystem and measuring all-by-all pairwise infections, it is possible to construct phage-bacteria interaction networks (PBINs) and analyze their underlying structure. Under controlled laboratory settings, phage-bacteria coevolution experiments rapidly generate rich ecological networks, allowing researchers to study the mechanisms underlying their formation (*7*, *18*–*20*). For example, arms race dynamics produce nested patterns when bacteria evolve escalating resistance and phages counter with increasing infectivity (*4*, *7*, *21*). Modules can form as a result of fluctuating selection dynamics (*18*, *20*, *22*) or the presence of different phage defense elements that confer specialized immunity (*23*). However, experiments have been unable to reproduce PBINs with multiscale structure (modular at broad scales and nested within modules) as observed in nature without artificially providing spatial isolation between evolving subpopulations (*24*). These findings reinforce the assumption that spatial structure and/or other external influences are a prerequisite for the evolution of intricate ecological networks. We decipher the molecular mechanisms of parasite-host coevolution and demonstrate that the process of coevolution itself is sufficient to drive the rapid emergence of complex multiscale networks in simple ecological contexts.

Coevolution drives multiscale networks

We began by inoculating three well-mixed cultures with isogenic populations of *Escherichia coli* K-12 (*25*) and a lytic strain of bacteriophage 21 (Φ21) (*26*). Cultures were incubated at 37°C; every 24 hours, 1% of each commu was propagated into new media, allowing ba ria and phages to coevolve for 21 days. Each day, we tested phage populations for receptor use expansion and found that in populations 2 and 3, phages evolved to use two new receptors (described below). As a preliminary test, we then isolated a single phage from each population and timepoint. By comparing phage receptor usage in population 2, we discovered that phages initially expanded and then contracted their receptor use, consistent with an initial arms race transitioning to more specialized interactions. Therefore, we focused our efforts on population 2 (replicate experiments demonstrated that this transition was repeatable, a topic revisited later in the manuscript).

To analyze the dynamics in population 2 we isolated coevolved phages and bacteria from communities every third day (74 phages and 128 bacteria in total) and measured all 9472 pairwise infections to generate a PBIN. Initially the network showed a nested pattern as bacteria and phages evolved broader resistance and host range over time (Fig. 1A and fig. S1, $P < 0.001$ for bacteria and phage, linear model). However, the nested pattern was disrupted between days 18 and 21. On day 21, 2 host isolates remained sensitive to phages from day 18. Phages isolated on day 21 could not infect any host isolates from the same day, although phage titer remained high (fig. S7). Notably, phage isolates from day 21 broadened their host range to infect day-18 hosts while losing the ability to infect day-15 hosts. This pattern of phages gaining the ability to infect recently isolated hosts while losing the ability to infect previously isolated hosts is inconsistent with conventional assumptions of host-range expansion through arms race dynamics leading to nested networks. Instead, we hypothesized that specialized interactions may have formed, creating modules in the network.

We tested for the emergence of modules using the LP-BRIM (label propagation and bipartite recursively induced modules) algorithm which searches for community configuration in a bipartite network that maximizes the modularity metric ($0 < Q_b \leq 1$, where 1 denotes full modularity; see Methods) (*27*–*29*). The LP-BRIM algorithm identified three modules (Fig. 1B; $Q_b = 0.210$, $P < 0.0001$; fig. S2B). The first module includes bacteria isolates from days 3 to 9 and, nearly exclusively, early phage isolates from days 3 to 12. By contrast, modules 2 and 3 include phage and bacteria isolates exclusively from days 12 to 21. Next, we tested all 3 modules for nestedness, using the "nestedness metric based on overlap and decreasing fill" (NODF) method (see Methods) (*30*); all 3 modules were significantly nested (module 1 NODF = 0.766, $P < 0.0001$; module 2 NODF = 0.802, $P < 0.0001$; module 3 NODF = 0.777, $P < 0.0001$). Notably, phage-bacteria

[1]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA. [2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA. [3]Department of Biology, University of Maryland, College Park, MD 20742, USA. [4]Department of Physics, University of Maryland, College Park, MD 20742, USA. [5]Institut de Biologie, École Normale Supérieure, 75005 Paris, France.
*Corresponding author. Email: jrmeyer@ucsd.edu
†Previous address: School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA.

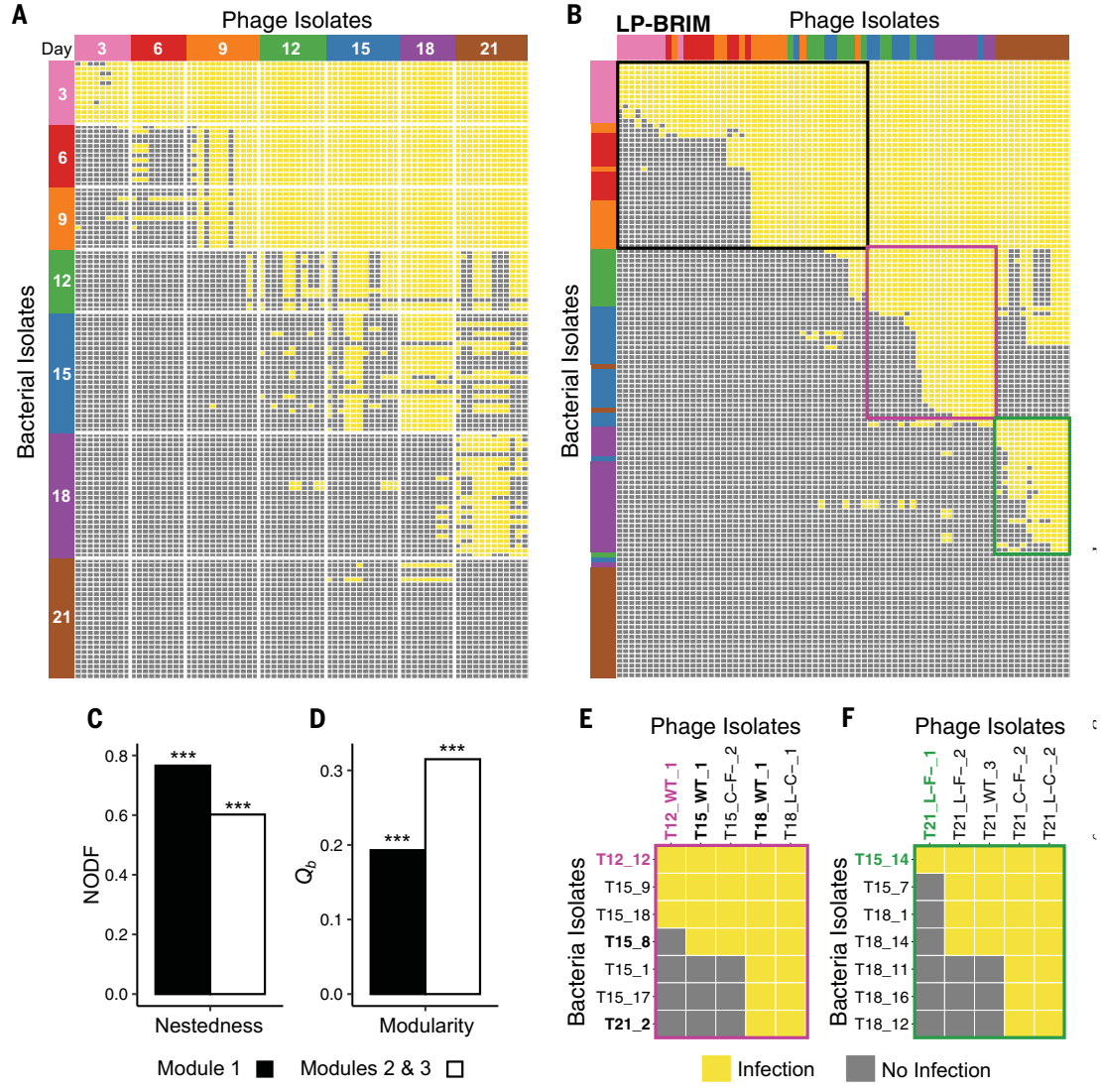


Fig. 1. Phage–bacteria coevolution generates multiscale network patterns in 21 days. (**A**) Phage–bacteria interaction network (PBIN) comprised of 9472 pairwise infections between 128 *E. coli* and 74 Φ21 strains isolated from various days of coevolution. (**B**) Community detection of the PBIN using the LP-BRIM algorithm reveals three modules (1, 2, and 3 indicated in black, pink, and green, respectively: Q_b = 0.2100***, n = 3). (**C**) Nestedness (NODF) and (**D**) Modularity (Q_b) of interactions between isolates from early (module 1, days 3 to 9: NODF = 0.7655, Q_b = 0.1935, n = 2) and late (modules 2 and 3, days 12 to 21: NODF = 0.6019, Q_b = 0.3151, n = 2) in the coevolution (P < 0.0001***). N refers to number of modules. P-values were obtained by comparing metrics (NODF or Q_b) against a null model where overall and marginal connectances were held fixed on average. (**E**) and (**F**) Infections between representative isolates from modules 2 and 3 were remeasured and recapitulated within-module nestedness. (E) and (F) serve as roadmaps for subsequent figures. Label aesthetics are applied consistently for continuity. PBINs that include isolate labels are available in the supplement (fig. S2, A and B).

infections amongst early isolates in module 1 were more nested (NODF = 0.766, P < 0.0001, Fig. 1C) than later isolates in modules 2 and 3 combined (NODF = 0.602, P < 0.0001, Fig. 1D) while infection between later isolates were significantly more modular (Q_b = 0.315, P < 0.0001) than early isolates (Q_b = 0.194, P < 0.0001)—a result of reduced cross-infections between phages of module 2 or 3 with bacteria of module 3 or 2, respectively.

To validate the nested pattern within modules as well as the repeatability of PBIN measurements, we remeasured representative interactions within modules 2 and 3, which recapitulated nestedness within each module (Fig. 1, E and F), and we independently remeasured 3654 interaction subsets of the full PBIN twice, which were significantly correlated with the full PBIN in Fig. 1A (P < 0.001, Mantel test, fig. S3). After demonstrating that coevolution between *E. coli* and phage Φ21 led to the emergence of a multiscale, nested-modular network, we endeavored to determine the evolutionary and molecular processes responsible for each pattern (i) initial nestedness, (ii) the formation of modules, and (iii) nestedness within modules.

Initial nestedness

We first investigated the factors driving the emergence of nestedness in module 1. Prior studies of coevolution between *E. coli* and phage λ show that arms race dynamics produce nested patterns as bacteria evolve resistance by mutating λ-receptor LamB and phages counter by innovating and using a new receptor, OmpF (*21*, *31*, *32*). We tested whether similar arms race dynamics drove nestedness in module 1 by measuring the frequency of host receptor mutations and phage receptor use in the population from days 0 to 12 (Fig. 2A). Population dynamics revealed multiple cycles of an arms race as bacteria sequentially mutated outer membrane proteins LamB, OmpC, and OmpF and phages that initially infected through LamB innovated to use OmpC and then OmpF. We then determined the receptor use of individual phage isolates by spotting them onto different lawns of bacteria with 2 of 3 relevant receptors deleted. This revealed that phages from day 3 could use either LamB or OmpC alone as receptors and phages from day 9 could use LamB, OmpC, or OmpF, representing the first instance of a documented triple-receptor phage we are aware of (Fig. 2B). Lastly, to confirm that consecutive receptor mutations and innovations were responsible for nestedness in module 1, we compared and found direct concordance between phage isolates from days 0, 3, and 9 infecting naturally coevolved bacteria and hosts genetically engineered with nonsense or deletion mutations in corresponding receptors (Fig. 2C). Altogether, our results confirm that multiple rounds of host receptor loss and phage receptor use gain drove initial nestedness in the network. By day 9 evolved bacteria had lost the first two receptors (LamB

and OmpC), and phages countered by evolving to use a third receptor, OmpF. We discovered that bacteria could evolve complete resistance by deleting the final OmpF receptor (day 12, 1 of 13 isolates), yet completely resistant mutants did not increase in frequency until day 21 (23 of 25 isolates). Bacterial fitness competitions revealed that the sequential loss of LamB, OmpC, and then OmpF receptors correlated with 14, 31, and 50% reductions in growth rate, respectively, with completely resistant isolates paying immense costs (fig. S4). These results support a coevolutionary mechanism by which bacteria maintained the OmpF receptor until phage-induced lysis became sufficiently intense to select for complete resistance between days 18 and 21.

Module formation

The separation of modules 2 and 3, comprised of phages and bacteria isolated on day 12 and beyond, implies that as phages gained the ability to infect new hosts, the phages lost infectivity on other, evolved hosts (see Fig. 1, B, E, and F, and fig. S2). Because bacteria retained the OmpF receptor and began evolving nucleotide substitutions within *ompF*, we hypothesized that modules formed as a result of specialized interactions between different phages and variants of OmpF. To investigate, we focused on bacteria and phages with the narrowest resistance or infectivity range within each module, as interactions between these isolates initiate the divergence of modules in the network. We refer to these strains as "entry" isolates. As expected, entry phages could infect within, but not between, modules 2 and 3 (Fig. 3A). Notably, this loss-for-gain pattern was recapitulated when phages were tested against ΔLamB ΔOmpC hosts engineered with *ompF* mutations from entry bacteria, confirming that interactions between different phages and OmpF variants drove module formation (Fig. 3A).

To illustrate this result, we mapped bacterial mutations onto solved structures of OmpF and phage mutations onto predicted structures of the host-recognition protein, (similar to λ central tail fiber protein, J) (Fig. 3B). We found that OmpF mutations were located on extracellular loops and J mutations were concentrated on extruding finger-like projections. The exposure of these sites on the outer surface of the host receptor and phage tail fiber strongly suggests that infectivity between entry phages and bacteria is governed by specific binding between accessible surfaces on variants of J and OmpF. Notably, although both the infectivity pattern between phages and OmpF variants is modular, the host and phage phylogenies show different shapes (figs. S5 and S6). The host phylogeny shows considerable diversity on days 15 and 18, spanning modules 2 and 3 (fig. S5), whereas the phage phylogeny is nested (fig. S6). Module 3 phage T21_L-F-_1 is a descendant of module 2 phage T12_WT_1 and has 7 additional *J* mutations, some or all of which cause gain of infectivity of module 3 OmpF variants and lost infectivity of module 2 OmpF variants. Phage mutations with the capacity to cause simultaneous gain and loss of hosts have been observed previously (*22*, *33*) and similar breakdowns between phylogenetic and phenotypic patterns have been observed in *E. coli*–λ coevolution (*21*).

Nestedness within modules

We investigated the third key pattern in our multiscale PBIN: the formation of nestedness within modules. In line with our observations for module formation, we hypothesized that additional *ompF* and *J* mutations produced nestedness within modules. Focusing on module 2 we used a similar approach as described above. First we measured interactions between representative phages and bacteria, recapitulating nestedness within the module (Fig. 4A). Notably, phylogenies of phage and bacterial isolates were also nested. As phages accumulated *J* mutations and bacteria accumulated *ompF* mutations, they gained increasing host and resistance range, respectively, although different *ompF* mutations conferred different amounts of protection (Fig. 4A). When we tested the same phages on ΔLamB ΔOmpC hosts engineered with *ompF* mutations from coevolved bacteria, we found the same pattern,

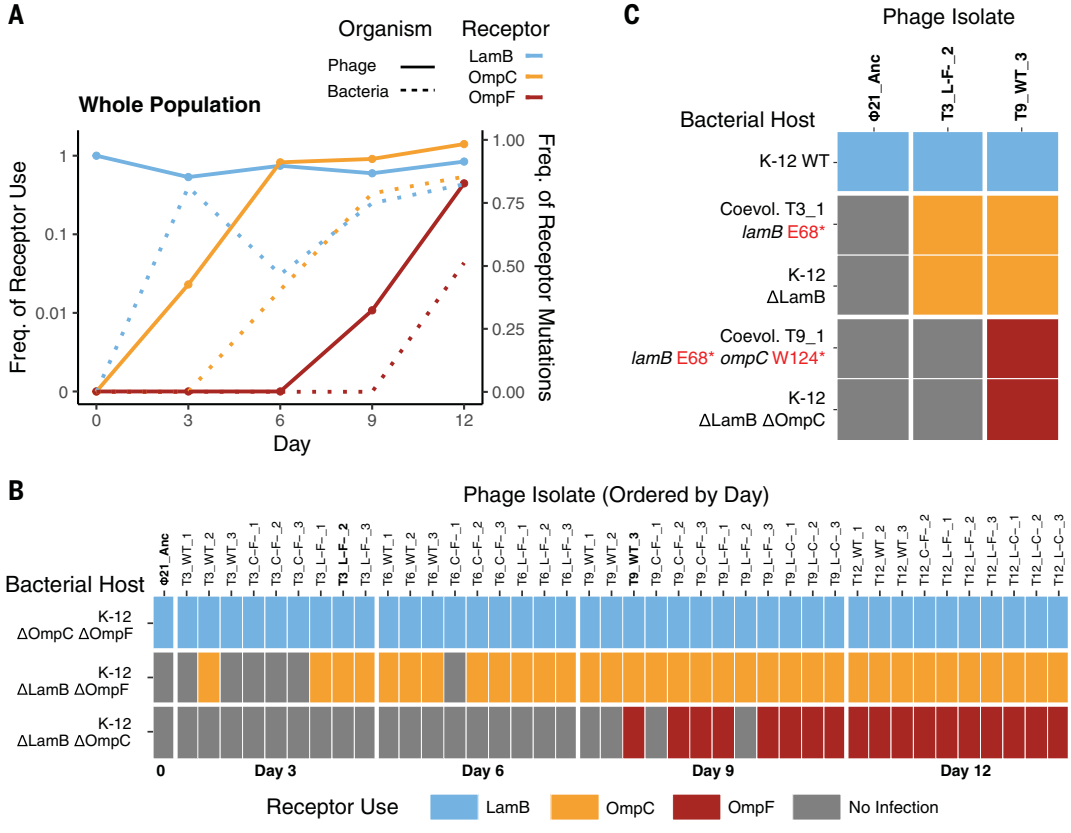


Fig. 2. Initial nestedness is driven by multiple cycles of host receptor loss and phage receptor innovation. (**A**) Frequency of phage receptor use (solid) and bacterial receptor mutations (dotted) in populations across coevolution days 0 to 12. Frequency of receptor use was calculated as the titer of phages on ΔOmpC ΔOmpF, ΔLamB ΔOmpF, or ΔLamB ΔOmpC K-12 hosts relative to K-12 WT. Frequency of receptor mutations was calculated as the frequency of whole-population sequencing reads with mutations affecting each receptor. (**B**) Ability of 41 phage isolates from coevolution days 0 to 12 to use host receptors LamB, OmpC, or OmpF, determined by spotting phages on agar infused with different dual-receptor knockout hosts. (**C**) Ability of phage isolates with expanding receptor use to infect naturally coevolved and genetically engineered bacteria with mutations successively eliminating receptors.

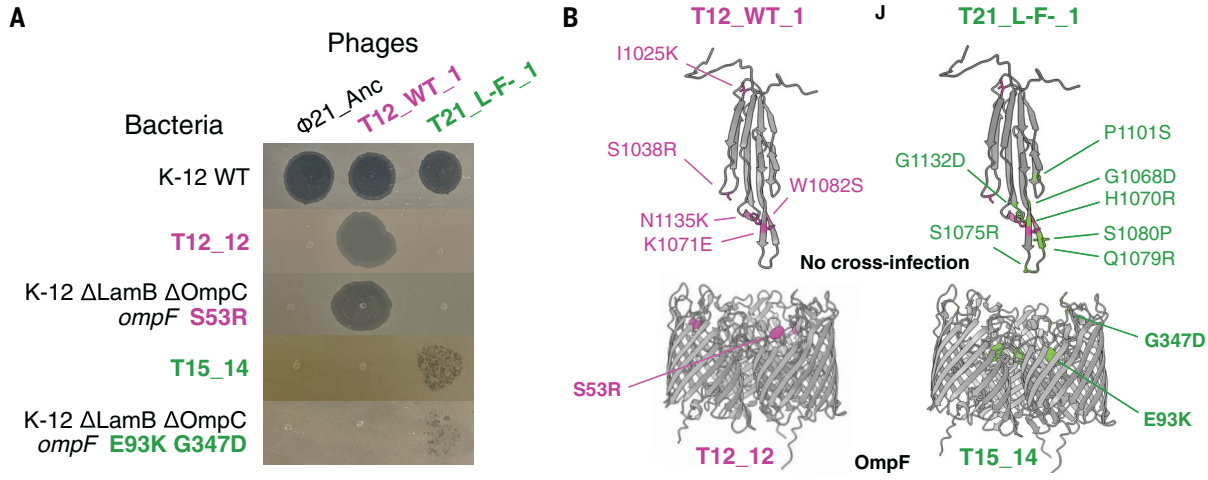


Fig. 3. Modules form as a result of specialized interactions between different phages and OmpF variants. (**A**) Infection assay of Φ21 and module entry phages spotted onto lawns of K-12 WT, module entry bacteria, and ΔLamB ΔOmpC hosts engineered with *ompF* mutations from corresponding entry bacteria. (**B**) AlphaFold predicted structures of the phage host-recognition central tail fiber protein J (top) and solved structure of host OmpF receptor (bottom). Mutations in module 1 and 2 isolates are annotated in pink and green, respectively. Phage T21_L-F-_1 has 7 mutations (green) in addition to the 5 mutations present in phage T12_WT_1 (pink). OmpF mutants do not share any mutations.

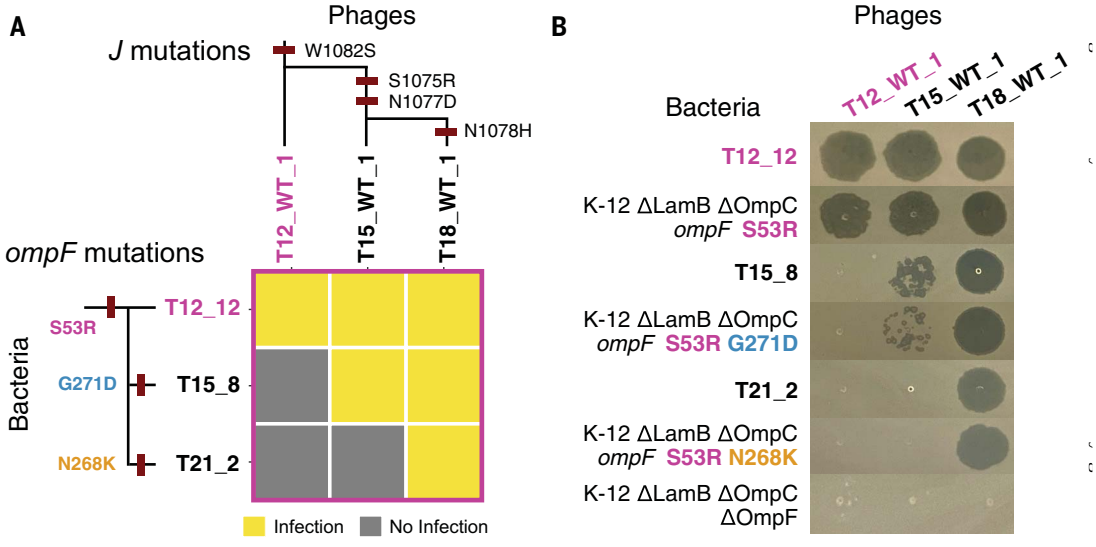


Fig. 4. Lock-and-key arms race dynamics create within-module nestedness. (**A**) Phylogenies of *J* and *ompF* mapped onto a PBIN of representative phage and bacterial isolates from module 2 (see Fig. 1E). Mutations are indicated on the phylogeny in red. OmpF mutations are labeled in pink, blue, and orange. Synonymous mutations and mutations in other genes are shown in figs. S5 and S6. (**B**) Infection assay of phages spotted onto lawns of coevolved bacteria or ΔLamB ΔOmpC hosts engineered with *ompF* mutations from corresponding bacteria.

confirming that within-module nestedness was driven by the accumulation of mutations in *J* and *ompF* (Fig. 4B).

Coexistence of modules over time

Our finding of a dynamic multiscale network raises the question of whether separate modules of interacting phages and bacteria, isolated at different timepoints, actually overlapped through time. We hypothesized that a multiscale infection network comprised of phage and bacterial isolates from modules 2 and 3 could have coexisted at single points in time, given two observations: (i) phage titers in day 21 flasks remained high even though bacterial isolates were resistant to all contemporary phage isolates (fig. S7 and Fig. 1A) and (ii) 2 of 25 day-21 bacterial isolates remained sensitive to module 2 phages from day 18. To test this we revived the frozen preserved community from day 21, passaged it for 3 additional days, and isolated bacteria and phages from day 24. We added the interactions of day-24 isolates to our network (fig. S8) and compared the host-range and resistance-range profiles of day-24 phages and bacteria to those of phage and bacterial isolates from modules 2 and 3 to determine whether progenitors from each module were present on the 24th day. Clustering analyses revealed that extant day-24 phages and bacteria each assorted into two major groups: one set clustered with module 2 and a second clustered with module 3 (fig. S9). Sequencing of representative day-24 phage isolates from each group (T24_C-F_1 and T24_L-C-_3) also confirmed that they were descendants of modules 2 and 3, respectively (fig. S6). Altogether, these results verify that even though modules 2 and 3 formed at separate times they coexisted and were maintained by ecological and coevolutionary processes occurring within the flask.

Discussion

In this study we demonstrate that coevolutionary processes are sufficient to generate complex multiscale ecological networks rapidly and without external influences or spatial structure. By determining the mechanistic basis of key interactions, we revealed how both punctuated and gradual evolutionary processes create three major patterns in the network (summarized in fig. S11). Initially nestedness emerges as bacteria and phages accumulate increasing resistance and infectivity through

multiple cycles of an arms race, whereby bacteria evolve mutations that functionally remove receptors and phages innovate to use new receptors. Modules form when phages evolve specialized interactions with mutated receptor variants, allowing them to gain infectivity of some hosts while losing infectivity of others. Finally, nestedness emerges within modules as phages and bacteria accumulate mutations to increase infectivity and resistance through lock-and-key arms race dynamics.

Although the present study focused on a single flask, the phenomena we discovered are repeatable. Many key features of the coevolution between *E. coli* and Φ21 were observed in replicate trials. For example, Φ21 evolved to use a second receptor, OmpC, in each of the 3 originally conducted replicates and 9 additional replicates conducted later in the same fashion. In 5 of 12 of these replicates, Φ21 evolved to use a third receptor, OmpF (fig. S10). In every case where OmpF use evolved, such use arose after phages evolved the ability to infect hosts through OmpC, suggesting similar mechanisms drove evolution within the replicate trials. Although we were able to characterize the molecular underpinnings of major features of the interaction network, we also note that more phenotypic and genomic variation exists which we have yet to explain, further highlighting the complexity that can evolve rapidly (i.e., within 100 to 200 generations). Although the molecular mechanisms identified in our flasks will differ across vastly different taxa and ecological communities (*24*, *34*), our work demonstrates how inherent coevolutionary processes can give rise to complex patterns that abound in natural ecosystems.

More broadly, previous observations of geographically isolated modules have suggested that spatial separation assists in the formation of such patterns (*5*, *8*, *9*, *24*). Although spatial structure certainly plays an important role in generating and maintaining diversity (*35*), it is also possible that modules may form in sympatry, and then separate in space over time. Similar pitfalls have been encountered when inferring past speciation processes from modern species distributions (*36*–*38*). Complementary to indirect inference, our finding provides a tractable experimental example of how diversity begets further diversity and, in turn, complex interaction structures (*39*, *40*) akin to Darwin's entangled bank (*41*). Although limits on diversity must exist, it seems that life—even in simple environments—is evolving far from these bounds.

REFERENCES AND NOTES

1. J. Bascompte, P. Jordano, C. J. Melián, J. M. Olesen, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9383–9387 (2003).
2. J. M. Olesen, J. Bascompte, Y. L. Dupont, P. Jordano, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19891–19896 (2007).
3. M. A. Fortuna *et al*., *J. Anim. Ecol.* **79**, 811–817 (2010).
4. C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, J. S. Weitz, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E288–E297 (2011).
5. S. Valverde *et al*., *Nat. Ecol. Evol.* **4**, 568–577 (2020).
6. D. Piel *et al*., *Nat. Microbiol.* **7**, 1075–1086 (2022).
7. J. Gurney *et al*., *Mol. Ecol.* **26**, 1764–1777 (2017).
8. C. O. Flores, S. Valverde, J. S. Weitz, *ISME J.* **7**, 520–532 (2013).
9. K. M. Kauffman *et al*., *Nat. Commun.* **13**, 372 (2022).
10. J. N. Thompson, *Am. Nat.* **153**, S1–S14 (1999).
11. B. Koskella, J. N. Thompson, G. M. Preston, A. Buckling, *Am. Nat.* **177**, 440–451 (2011).
12. O. Eriksson, *Biol. Rev.* **91**, 168–186 (2014).
13. S. T. Segar *et al*., *Trends Ecol. Evol.* **35**, 454–466 (2020).
14. P. R. Guimarães Jr., M. M. Pires, P. Jordano, J. Bascompte, J. N. Thompson, *Nature* **550**, 511–514 (2017).
15. S. J. Beckett, H. T. P. Williams, *Interface Focus* **3**, 20130033 (2013).
16. S. E. Forde *et al*., *Nature* **455**, 220–223 (2008).
17. J. S. Weitz *et al*., *Trends Microbiol.* **21**, 82–91 (2013).
18. A. R. Hall, P. D. Scanlan, A. D. Morgan, A. Buckling, *Ecol. Lett.* **14**, 635–642 (2011).
19. M. A. Fortuna *et al*., *Evolution* **73**, 1001–1011 (2019).
20. K. E. Kortright, B. K. Chan, B. R. Evans, P. E. Turner, *J. Evol. Biol.* **35**, 1475–1487 (2022).
21. A. Gupta *et al*., *Ecol. Lett.* **25**, 876–888 (2022).
22. P. Gómez, A. Buckling, *Science* **332**, 106–109 (2011).
23. F. A. Hussain *et al*., *Science* **374**, 488–492 (2021).
24. E. Shaer Tamar, R. Kishony, *Nat. Commun.* **13**, 7971 (2022).
25. T. Baba *et al*., *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
26. M. Feiss *et al*., *Microbiol. Resour. Announc.* **11**, e0012022 (2022).
27. M. J. Barber, *Phys. Rev. E* **76**, 066102 (2007).
28. C. O. Flores, T. Poisot, S. Valverde, J. S. Weitz, *Methods Ecol. Evol.* **7**, 127–132 (2016).
29. X. Liu, T. Murata, *Trans. Jpn. Soc. Artif. Intell.* **25**, 16–24 (2010).
30. M. Almeida-Neto, P. Guimarães, P. R. Guimarães Jr., R. D. Loyola, W. Ulrich, *Oikos* **117**, 1227–1239 (2008).
31. J. R. Meyer *et al*., *Science* **335**, 428–432 (2012).
32. A. Gupta *et al*., *eLife* **11**, e76162 (2022).
33. S. Duffy, C. L. Burch, P. E. Turner, *Evolution* **61**, 2614–2622 (2007).
34. H. Georjon, A. Bernheim, *Nat. Rev. Microbiol.* **21**, 686–700 (2023).
35. J. N. Thompson, *The Geographic Mosaic of Coevolution* (Univ. of Chicago Press, 2005).
36. M. Doebeli, U. Dieckmann, *Am. Nat.* **156** (S4), S77–S101 (2000).
37. M. Doebeli, U. Dieckmann, *Nature* **421**, 259–264 (2003).
38. E. J. Richards, J. W. Poelstra, C. H. Martin, *Evol. Lett.* **2**, 524–540 (2018).
39. N. Madi, M. Vos, C. L. Murall, P. Legendre, B. J. Shapiro, *eLife* **9**, e58999 (2020).
40. A. A. Forbes, T. H. Q. Powell, L. L. Stelinski, J. J. Smith, J. L. Feder, *Science* **323**, 776–779 (2009).
41. C. Darwin, On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life (John Murray, 1859).
42. J. Borin, joshborin/phi21multiscalePBIN: Data repository for multiscale pbin paper, Zenodo (2023); https://zenodo.org/records/8277766.
43. A. Lucia-Sanz, aluciasanz/nestedness_modularity_pbin_analysis, v1.1, Zenodo (2023); https://zenodo.org/records/8277861.

ACKNOWLEDGMENTS

We thank S. Casjens and the R. Young Laboratory for providing bacteriophage 21 (Φ21) and its annotated genome. We thank S. Ardell, the Kryazhimskiy Lab, and other members of the Meyer Lab for useful discussions and comments on the manuscript. **Funding:** This work was supported by the following: Howard Hughes Medical Institute Emerging Pathogens Initiative grant 7012574 (J.R.M.) National Institutes of Health grant T32GM007240 (J.M.B.) Army Research Office grant W911NF1910384 (J.S.W.) Chaires Blaise Pascal program of the Île-de-France region (J.S.W.). **Author contributions:** Conceptualization: J.M.B. and J.R.M. Methodology: J.M.B., J.J.L., A.L.S., and K.R.G. Investigation: J.M.B., J.J.L., A.L.S., K.R.G., J.S.W., and J.R.M. Visualization: J.M.B., A.L.S., and K.R.G. Supervision: J.S.W. and J.R.M. Writing – original draft: J.M.B. and J.R.M. Writing – review and editing: J.M.B., J.J.L., A.L.S., K.R.G., J.S.W., and J.R.M. **Competing interests:** Authors declare that they have no competing interests. **Data and materials availability:** Data and code are available online (*42*), as well as code for network analyses (*43*). Sequences included in this study have been deposited in the NCBI Sequence Read Archive under BioProject accession number PRJNA1006779. Biological material is available from J.R.M. under the material transfer agreement with UC San Diego (https://blink.ucsd.edu/research/conducting-research/mta/index.html). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi5536
Materials and Methods
Figs. S1 to S11
Tables S1 to S4
References (*44*–*54*)
MDAR Reproducibility Checklist

Submitted 3 May 2023; accepted 28 September 2023
10.1126/science.adi5536

HERBIVORY

Plant size, latitude, and phylogeny explain within-population variability in herbivory

The Herbivory Variability Network*†

Interactions between plants and herbivores are central in most ecosystems, but their strength is highly variable. The amount of variability within a system is thought to influence most aspects of plant-herbivore biology, from ecological stability to plant defense evolution. Our understanding of what influences variability, however, is limited by sparse data. We collected standardized surveys of herbivory for 503 plant species at 790 sites across 116° of latitude. With these data, we show that within-population variability in herbivory increases with latitude, decreases with plant size, and is phylogenetically structured. Differences in the magnitude of variability are thus central to how plant-herbivore biology varies across macroscale gradients. We argue that increased focus on interaction variability will advance understanding of patterns of life on Earth.

Plant-herbivore interactions, which involve more than half of macroscopic biodiversity and 90% of macroscopic biomass (*1*), are believed to shape macroscale biological patterns and processes, such as plant and herbivore biodiversity gradients, biomass distributions, community structure, species coexistence, and trait evolution (*2*–*4*). Biologists have studied the role of herbivory at macroscales by quantifying how the mean herbivore damage level covaries with latitude, biome, functional traits, and phylogeny (*5*–*7*). However, macroscale patterns have not always matched expectations. For example, despite the paradigm that herbivore pressure increases toward the equator owing to more-benign environmental conditions, empirical patterns have been weak or inconsistent (*8*–*10*). Similarly, despite the expectation that closely related plant species should face similar pressures from herbivores, phylogenetic signal in mean herbivore damage is often undetectable or restricted to certain groups (*5*, *11*). We suggest that our understanding of macroscale patterns in herbivory can be improved by considering patterns in the magnitude of variability in herbivory rather than only mean interaction strength.

Variability is a hallmark of plant-herbivore interactions (*12*). Within populations, patterns in damage are often highly skewed, with most plant individuals receiving very low levels of damage, and a few plants receiving high levels (*13*). Although there are limited data on the drivers and consequences of this variability, theory indicates that within-species variation in traits or interactions can be as important as the mean for biological processes ranging from population viability to evolutionary dynamics (*14*, *15*). For example, spatial variability can stabilize plant-herbivore dynamics by giving plants refuges from overexploitation (*16*), can increase the importance of competition among herbivores (*17*), can maintain diversity by facilitating the evolutionary coexistence of alternative strategies (*18*), and can drive disease dynamics by causing superspreading events (*19*). Variation in damage among plant individuals also indicates the potential pattern of selection by herbivores, which drives plant defense evolution (*20*). Variability has been hypothesized to favor inducible plant defenses over constitutively expressed defenses—a central dichotomy in defense evolution (*21*). Despite the central role that variability likely plays in the ecology and evolution of plants and herbivores, macroscale patterns of variability remain uncharacterized. In this work, we propose and test three hypotheses for patterns in the magnitude of variation in herbivore damage among individuals within plant populations.

First, we hypothesize that herbivory variability within populations increases with distance from the equator owing to shorter growing seasons and less-stable abiotic conditions at higher latitudes reducing the time available for herbivore foraging. A latitudinal variability gradient could help explain how herbivores have influenced global patterns of plant biodiversity despite the weak latitudinal gradient in mean herbivory (*22*, *23*). Herbivory may maintain plant diversity at low latitudes not just by being more intense on average but by being a more-consistently important force within plant populations. Second, we hypothesize that herbivory is more variable among small plants compared with large plants. Large

*Corresponding author. Email: william.wetzel@montana.edu
†Herbivory Variability Network authors and affiliations are listed in the supplementary materials.

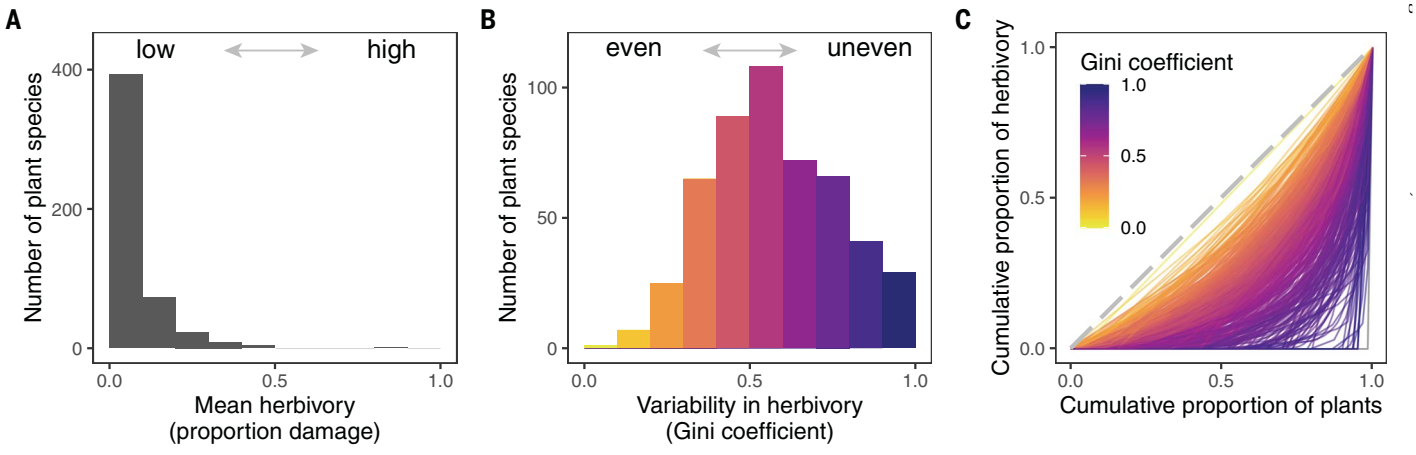


Fig. 1. Mean and variability in plant-herbivore interactions. (**A**) Histogram of the number of plant species with different mean proportion leaf area damaged by herbivores. (**B**) Histogram of the Gini coefficient values for all plant species in our dataset. (**C**) Lorenz curves from all 790 population surveys in our dataset. Each curve shows the cumulative proportion of herbivory across the cumulative proportion of plants, ordered by increasing herbivory, for one plant population. Curves closer to the 1:1 line (gray dashes) indicate more-even distributions. Lorenz curves form the basis for the calculation of the Gini coefficient of inequality, which ranges from 0 (a perfectly even distribution) to 1 (a perfectly uneven distribution). Curves are colored by their Gini coefficient [as in (B)]. Sample sizes are 790 surveys of 503 plant species.

plants, which represent a greater sampling area, should average over small-scale random variation in herbivory, resulting in values closer to the population mean, whereas small plants should be more likely to escape herbivory entirely or be highly damaged by a few events. If supported, this hypothesis would expand our understanding of long-studied differences in defenses between trees and herbs (*24*), with consistent damage on large plants explaining why trees invest a greater proportion of their biomass in constitutive defenses (*25*). Third, we hypothesize that variability in herbivory is phylogenetically structured, with more-closely related plants displaying more-similar levels of variability. This pattern, which has been documented for mean herbivory (*5*), would indicate that variability is influenced by species-level traits and is not simply random, as it has often been treated.

To characterize macroscale patterns in population-level mean and variability in herbivory, 127 research teams in 34 countries used a standardized protocol (*26*) to sample plants and quantify aboveground herbivore damage for 790 populations of 503 species in 135 families. This sample comprised more than 50,000

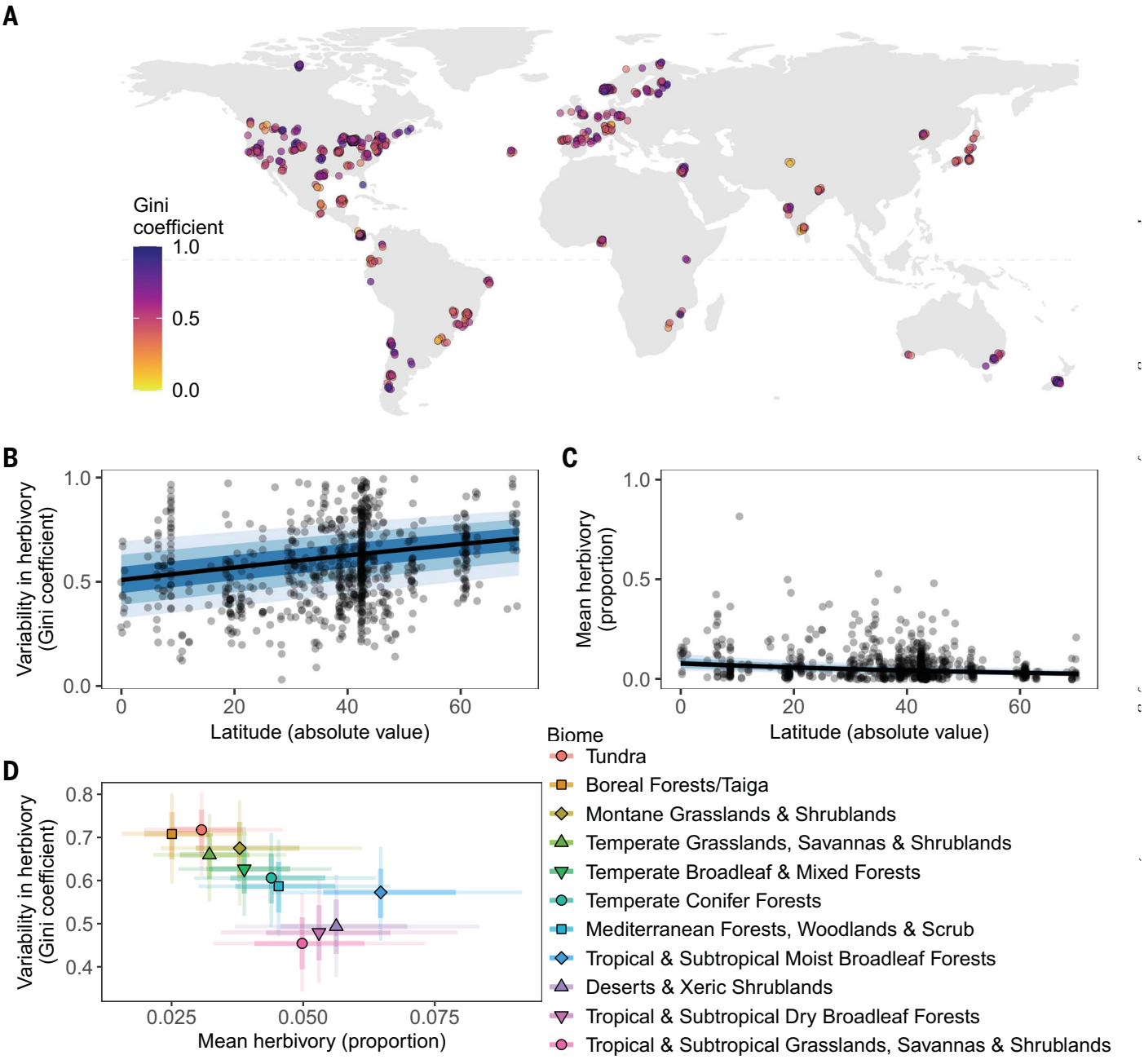


Fig. 2. Global patterns of variability in herbivory within plant populations. (**A**) The geographic distribution of our sampling sites, colored by variability in herbivory among individuals within populations (Gini coefficient). Points are slightly jittered for visibility. (**B** and **C**) Variability in herbivory increased (B) and mean herbivory decreased (C) with latitude across our sampling extent. Lines show predicted means and 50, 80, and 95% credible intervals from Bayesian phylogenetic beta regressions. (**D**) The 11 biomes in our study can be characterized by their mean and variability in herbivory. Herbivory variability and mean showed an inverse relationship across biomes [ρ = −0.67 (−0.94 to −0.08)], but there were also differences in variability between biomes with similar means. Error bars show 50 and 80% credible regions. Sample size is 790 surveys of 503 species. Legend in (D) is ordered by Gini coefficient.

plant individuals distributed across six continents and 116° of latitude. Past macroscale studies that have focused on differences in means typically examined relatively few individuals per population (*5*). By contrast, we sampled 60 individuals per population, which allowed us to analyze patterns in population-level variability. For each plant individual, we recorded plant size (height for most species, or canopy diameter for prostrate species) and visually estimated the cumulative proportion of leaf tissue damaged by invertebrate and vertebrate herbivores. We quantified the variability in herbivory among individuals within populations using the Gini coefficient—a commonly used scale-invariant metric that ranges from 0 to 1 (perfectly even to perfectly uneven) (*27*). We tested our hypotheses by quantifying associations between each macroscale factor and the Gini coefficient or mean herbivory using Bayesian phylogenetic beta regressions.

Overall, within-population variation in herbivore damage was very high [mean Gini coefficient = 0.61 (95% confidence interval: 0.40 to 0.78)] (Fig. 1). On average, the most-damaged individual in each plant population lost 34.2% (32.4 to 36.0%) of its leaf area to herbivory, whereas 27.9% (25.9 to 29.9%) of individuals completely or essentially escaped herbivory (<0.5% damage). Half of the damage in each population was concentrated on 11.3% (10.7 to 11.9%) of its individuals on average. The level of variation within populations also varied significantly across populations and species, with the Gini coefficient ranging from 0.03, an almost perfectly even distribution of damage, to 1.0, a perfectly uneven distribution with all damage on one plant (Fig. 1, B and C). Even though the Gini coefficient normalizes by the mean, it can nevertheless be correlated with it. Mean herbivory and the Gini coefficient were negatively correlated, with Gini coefficients being low for the 3.9% of populations with very high (>25%) mean herbivory, whereas populations with lower mean herbivory exhibited the full range of Gini coefficients ($\rho = -0.46$) (fig. S1).

Geographic patterns of variability

We found strong support for the latitudinal variability gradient hypothesis (Fig. 2, A and B). Variation was lowest at the equator [Gini = 0.51 (0.33 to 0.69)] and increased toward 70°N and 70°S (70°N/S) [Gini = 0.70 (0.54 to 0.84); Bayesian coefficient of determination (R^2) = 5%; posterior probability (p_p) = 1.0; Bayes factor (BF) = 2.0×10^4]. Mean herbivory, by contrast, declined with latitude, from 8.0% (4.1 to 12.3%) at the equator to 2.9% (1.4 to 4.7%) at 70°N/S; this relationship was less predictable than the one for the Gini coefficient (R^2 = 2%; p_p = 1.0; BF = 2.9×10^4) (Fig. 2C, figs. S2 and S3, and tables S1 to S3). Thus, plants at higher latitudes, with shorter growing seasons and lower temperatures (*26*), receive less herbivory on average, and that herbivory is concentrated on fewer individuals. This result could conceivably be an artifact of the negative mean–Gini coefficient correlation. We therefore repeated our analysis with mean herbivory included as a covariate. The estimated latitudinal variability gradient was still strongly positive, though it was lower in magnitude, with a 20% (6 to 38%) increase in the Gini coefficient from the equator to 70° N/S (R^2 = 23%; p_p = 1.0; BF = 14.5) (fig. S4). This relationship captured differences among biomes: Higher latitude and higher elevation biomes had higher Gini coefficients and lower mean herbivory (Fig. 2D and fig. S5). Whereas there was a negative correlation between the mean and Gini coefficient among biomes [$\rho = -0.68$ (-0.95 to -0.10)], there were also large differences in the Gini coefficient between biomes with similar mean herbivory. This suggests that interaction variability could be a fundamental characteristic differentiating biological systems across macroscales.

Debate over the contribution of herbivory to global patterns of plant evolution has been contentious (*3*, *6*, *8*, *10*, *22*, *23*). Our data show strong evidence of a meaningful, although noisy, latitudinal decline in mean levels of herbivore damage. They also show that herbivory becomes more variable with increasing latitude. This pattern is consistent with our hypothesis that herbivory influences plant evolution at low latitudes not just by being more intense on average, but also by being more consistently important within a plant population. Theory predicts that the relationship between the strength of antagonistic interactions and the intensity of selection is concave down (saturating) at low mean interaction strengths (*28*), which means that variability at high latitudes, where mean herbivory is low, should erode selection through nonlinear averaging (*14*), all else being equal. Our finding is also consistent with the hypothesis that inducible defenses are more common among temperate compared with tropical plants (*29*, *30*) because greater variation in herbivory is predicted to select for inducibility (*21*). In addition to seasonality and climate, other mechanisms for the latitudinal variability gradient could include greater predation pressure on herbivores at low latitudes (*3*) suppressing localized outbreaks and high tropical herbivore diversity and specialization (*31*) evening out damage patterns across plant individuals. More generally, our results confirm the long-held view that biotic interactions are more consistent in the tropics, perhaps owing to longer growing seasons or greater species diversity and specialization (*3*).

Variability and plant size

We also found strong support for the size-mediated variability hypothesis. Populations of larger individuals exhibit less variability in herbivory among individuals. A 2-m increase in mean plant size (from 0.05 to 2.05 m, encompassing ~90% of our populations) resulted in a 32.7% (20.6 to 44.7%) decrease in the Gini coefficient [from 0.70 (0.54 to 0.85) to 0.47 (0.29 to 0.66); R^2 = 13.3%; p_p = 1.0; BF = 4.6×10^7] (Fig. 3A and fig. S6). This relationship held even after accounting for the decline in plant size with increasing latitude and differences in plant abundance (which ranged from 2 to 100% cover in our dataset) (tables S4 and S5) (*32*). Woody species, which averaged 4.1 times as large as herbs in our dataset, had 10.9% (2.9 to 19.1%) lower Gini coefficients compared with herbaceous species [0.56 (0.37 to 0.76) versus 0.63 (0.44 to 0.81); BF = 4.25]. However, the overall variance explained by growth form, including climber and graminoid categories, was low (R^2 = 2.8%) (Fig. 3B and fig. S7), which suggests that mean size is a more-important determinant of herbivory patterns than growth form. Mean herbivory, by contrast,

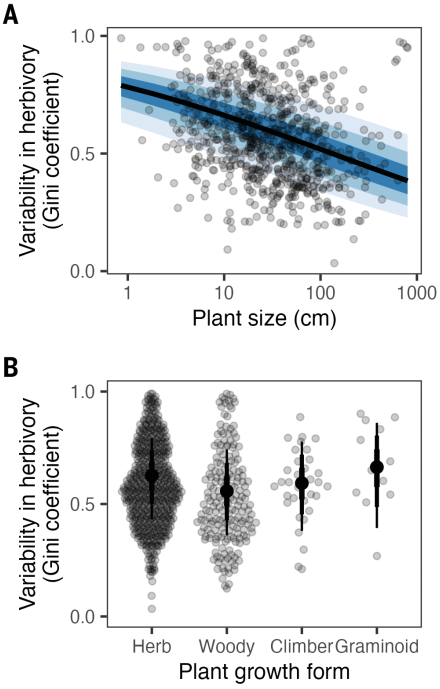


Fig. 3. Plant size shapes variability in herbivory. (**A**) Variability in herbivory among individuals within populations declines with the average size (height or canopy diameter for prostrate species) of plants in the population (R^2 = 13.3%; p_p = 1.0; BF = 4.6×10^7; 735 surveys of 472 species). (**B**) Variability in herbivory, however, is only weakly related to plant growth form (R^2 = 2.8%), with woody plants having 10.9% (2.9 to 19.1%) lower Gini coefficients compared with herbaceous species (790 surveys of 503 species). Lines, shaded regions, and large points show predicted means and 50, 80, and 95% credible intervals from phylogenetic Bayesian beta regressions. Each small gray point is one survey.

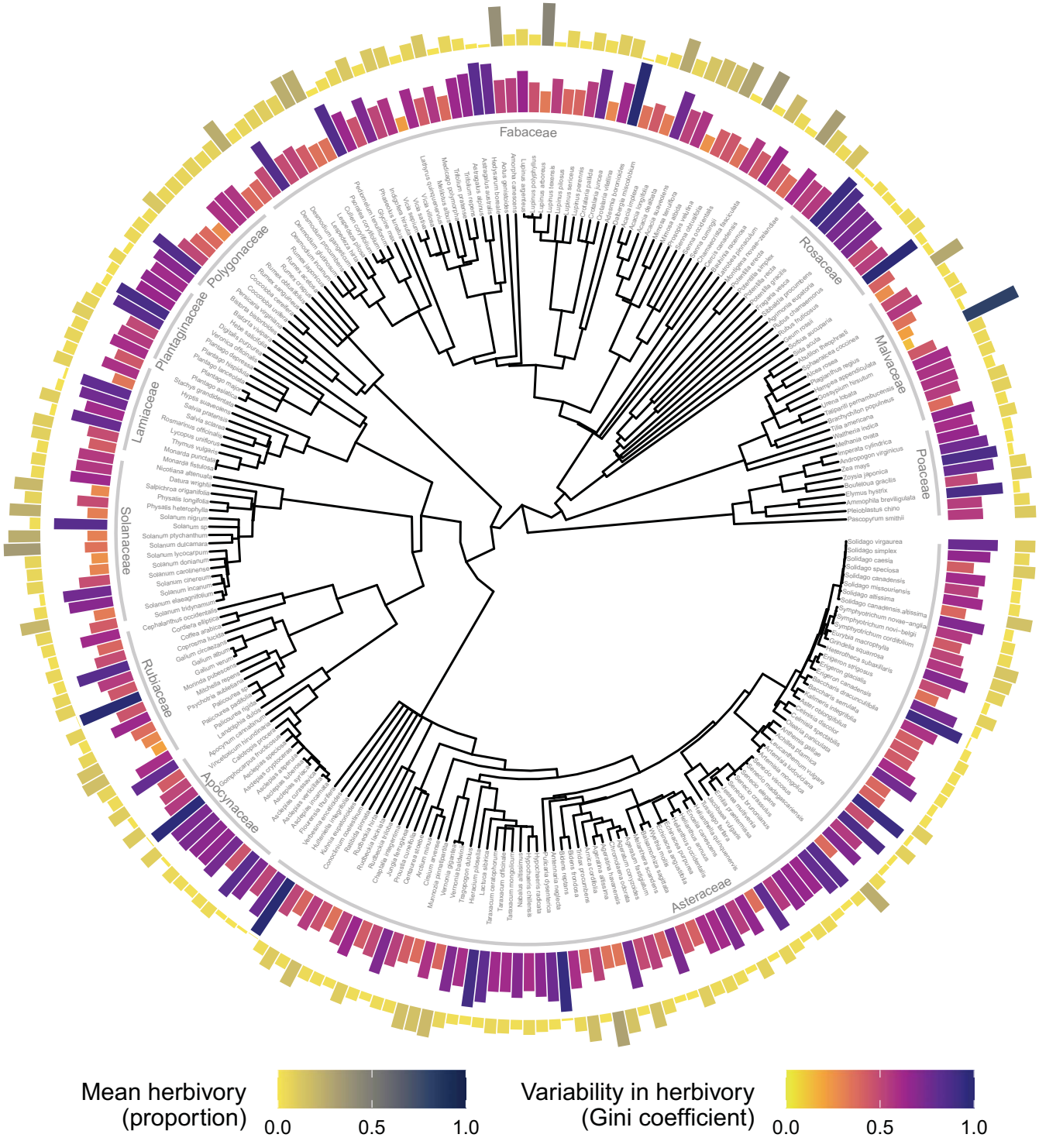


Fig. 4. Phylogenetic patterns of mean and variability in herbivory. Variability in herbivory among plants within populations (Gini coefficient) show greater phylogenetic signal [Pagel's λ = 0.51 (0.45 to 0.52); $P < 0.001$] compared with mean herbivory levels [Pagel's λ = 0.07 (0.06 to 0.08); $P > 0.1$]. For clarity, this tree includes only the 240 species from the 11 best-represented plant families (≥8 species per family). Our analyses included all 503 species in the dataset (see fig. S10 for the full tree).

was unrelated to mean size or growth form (figs. S8 and S9).

We posit that lower among-individual variability in herbivory on large plants results from the law of large numbers, which says that processes that involve more random events produce values closer to the overall mean. In other words, large plants, which have a greater number of potential herbivory events, average over within-plant variability and receive values closer to the population mean on average. Small plants, by contrast, are more likely to escape herbivory entirely or be severely damaged by a few events, which results in high variability. A key implication of this phenomenon is that

larger species (and larger stages within species) should experience greater selection for high concentrations of constitutive defenses or tolerance. Smaller species (and stages), by contrast, should experience greater selection for inducible defenses and low concentrations of metabolically cheap toxins to save resources in the absence of herbivory and repel herbivores when encountered. This dichotomy in defense evolution has been the focus of decades of research on differences in defenses between trees and herbs (*24*) and across ontogenetic stages (*33*). Whereas previous work has invoked complex biological explanations for these differences—such as how apparent plants are to herbivores (*24*)—our results suggest that patterns are more parsimoniously explained by the statistical consequences of mean plant size.

Phylogenetic patterns of variability

Finally, we tested the hypothesis that variability in herbivory is phylogenetically structured. The Gini coefficient exhibited significant phylogenetic signal [Pagel's $\lambda = 0.51$ (0.45 to 0.52); $P < 0.001$], indicating that more-closely related species display more-similar variability levels (Fig. 4 and fig. S10). Mean herbivory, by contrast, did not show meaningful phylogenetic signal [$\lambda = 0.07$ (0.06 to 0.08); $P = 1.0$]. These results were robust to tree topology and species sampling (supplementary materials). Our findings suggest that the mean damage level across species changes relatively rapidly in response to evolutionarily labile plant traits, whereas the variability is more strongly determined by traits that are phylogenetically conserved. Traits thought to influence the amount of herbivore damage, such as chemical defenses, diverge as plants escape their herbivores by evolving novel defenses (*2*, *34*), whereas characteristics such as geographic location and plant size, which we find relate to variability, tend to be less labile. High variability in some families (e.g., Apocynaceae and Plantaginaceae) invites further investigation and could help reveal drivers of these conserved patterns. To examine macroevolutionary patterns, we fit Brownian motion and Ornstein-Uhlenbeck models to test for differences in rates of evolution and the strength of stabilizing selection. The best-fitting models included optima for variability and mean herbivory in tropical versus temperate systems and woody versus herbaceous growth forms (tables S6 and S7), which indicates that the evolution of variability in herbivory seems to be driven by conserved plant traits and is therefore a biologically informative feature rather than random noise.

Conclusions

The assumption that plant-herbivore interactions are highly variable has long dominated ecology and evolution, with foundational works on so-called variable plants and herbivores (*12*) and theory exploring the consequences of variable herbivory (*21*). Our data confirm this assumption but also reveal a pattern that had not been previously documented: strong differentiation across systems in the level of variability itself. Variation in herbivory covaried with factors central to the ecology and evolution of plant-herbivore interactions, such as latitude, biome, plant size, and phylogeny. These macroscale patterns were often stronger than patterns for mean herbivory levels. This suggests that the level of variability could be important for driving differences in plant-herbivore biology around the planet, between species with different traits and across phylogeny. Although the importance of variability in interactions has been recognized by a few fields, such as epidemiology (*19*), the central role of interaction variability in shaping macroscale patterns of life on Earth has been underappreciated. Our global dataset is evidence for the ubiquity and predictability of variability in one biotic interaction and highlights the promise of further explorations of the causes and consequences of interaction variability.

REFERENCES AND NOTES

1. Y. M. Bar-On, R. Phillips, R. Milo, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6506–6511 (2018).
2. P. R. Ehrlich, P. H. Raven, *Evolution* **18**, 586–608 (1964).
3. D. W. Schemske, G. G. Mittelbach, H. V. Cornell, J. M. Sobel, K. Roy, *Annu. Rev. Ecol. Evol. Syst.* **40**, 245–269 (2009).
4. O. J. Schmitz, *Annu. Rev. Ecol. Evol. Syst.* **39**, 133–152 (2008).
5. M. M. Turcotte, T. J. Davies, C. J. M. Thomsen, M. T. J. Johnson, *Proc. R. Soc. B* **281**, 20140555 (2014).
6. A. T. Moles *et al*., *New Phytol.* **191**, 777–788 (2011).
7. S. Rasmann, A. A. Agrawal, *Ecol. Lett.* **14**, 476–483 (2011).
8. A. T. Moles, S. P. Bonser, A. G. B. Poore, I. R. Wallis, W. J. Foley, *Funct. Ecol.* **25**, 380–388 (2011).
9. D. Salazar, R. J. Marquis, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12616–12620 (2012).
10. J. Y. Lim, P. V. A. Fine, G. G. Mittelbach, *Glob. Ecol. Biogeogr.* **24**, 1106–1112 (2015).
11. X. Moreira *et al*., *Phytochemistry* **153**, 64–73 (2018).
12. R. F. Denno, M. S. McClure, in *Variable Plants and Herbivores in Natural and Managed Systems*, R. F. Denno, M. S. McClure, Eds. (Academic Press, 1983), pp. 1–12.
13. W. C. Wetzel, B. D. Inouye, P. G. Hahn, S. R. Whitehead, N. Underwood, *Annu. Rev. Ecol. Evol. Syst.* 10.1146/annurev-ecolsys-102221-045015 (2023).
14. D. I. Bolnick *et al*., *Trends Ecol. Evol.* **26**, 183–192 (2011).
15. C. Violle *et al*., *Trends Ecol. Evol.* **27**, 244–252 (2012).
16. R. M. Anderson, R. M. May, *J. Anim. Ecol.* **47**, 219–247 (1978).
17. R. F. Denno, M. S. McClure, J. R. Ott, *Annu. Rev. Entomol.* **40**, 297–331 (1995).
18. A. A. Agrawal, J. A. Lau, P. A. Hambäck, *Q. Rev. Biol.* **81**, 349–376 (2006).
19. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, *Nature* **438**, 355–359 (2005).
20. A. A. Agrawal, A. P. Hastings, M. T. J. Johnson, J. L. Maron, J.-P. Salminen, *Science* **338**, 113–116 (2012).
21. F. Adler, R. Karban, *Am. Nat.* **144**, 813–832 (1994).
22. A. T. Moles, J. Ollerton, *Biotropica* **48**, 141–145 (2016).
23. D. N. Anstett, K. A. Nunes, C. Baskett, P. M. Kotanen, *Trends Ecol. Evol.* **31**, 789–802 (2016).
24. P. Feeny, in *Biochemical Interaction Between Plants and Insects*, J. W. Wallace, R. L. Mansell, Eds. (Springer, 1976), pp. 1–40.
25. A. M. Smilanich, R. M. Fincher, L. A. Dyer, *New Phytol.* **210**, 1044–1057 (2016).
26. See the supplementary materials and methods.
27. J. L. Gastwirth, *Rev. Econ. Stat.* **54**, 306–316 (1972).
28. C. W. Benkman, *Ecol. Lett.* **16**, 1054–1060 (2013).
29. R. J. Bixenmann, P. D. Coley, A. Weinhold, T. A. Kursar, *Ecol. Evol.* **6**, 6037–6049 (2016).
30. P. D. Coley, M.-J. Endara, T. A. Kursar, *Oecologia* **187**, 361–376 (2018).
31. L. A. Dyer *et al*., *Nature* **448**, 696–699 (2007).
32. A. T. Moles *et al*., *J. Ecol.* **97**, 923–932 (2009).
33. K. Boge, R. J. Marquis, *Trends Ecol. Evol.* **20**, 441–448 (2005).
34. T. A. Kursar *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18073–18078 (2009).
35. W. Wetzel *et al*., Data for: Plant size, latitude, and phylogeny explain within-population variability in herbivory, data set, Dryad (2023); https://doi.org/10.5061/dryad.44j0zpckm.
36. W. C. Wetzel, HerbVar-Network/HV-Large-Patterns-MS-public: v1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.8133118.

ACKNOWLEDGMENTS

The authors thank A. Agrawal, J. Thaler, J. Conner, J. Lill, M. Weiss, N. Sanders, and M. Meek for helpful discussions and comments on different versions of the analyses and manuscripts. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US government. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official US Department of Agriculture (USDA) or US government determination or policy. **Funding:** The authors acknowledge funding for central project coordination from NSF Research Coordination Network grant DEB-2203582; the Ecology, Evolution, and Behavior Program at Michigan State University; and AgBioResearch at Michigan State University. Site-specific funding is listed in the supplementary materials. **Author contributions:** This project was conceptualized by W. C. Wetzel and N. Underwood. It was coordinated by W. C. Wetzel, M. L. Robinson, P. G. Hahn, B. D. Inouye, N. Underwood, S. R. Whitehead, K. C. Abbott, E. M. Bruna, N. I. Cacho, and L. A. Dyer. Other author contributions are listed in the supplementary materials. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The dataset generated and analyzed in this study is available at Dryad (*35*). Our code is archived at Zenodo (*36*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adh8830
Herbivory Variability Network Authors
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S7
References (*37*–*58*)
MDAR Reproducibility Checklist

Submitted 26 March 2023; accepted 27 September 2023
10.1126/science.adh8830

RADIATIVE COOLING

A solution-processed radiative cooling glass

Xinpeng Zhao[1]†, Tangyuan Li[1]†, Hua Xie[1]†, He Liu[1]†, Lingzhe Wang[2], Yurui Qu[3], Stephanie C. Li[1], Shufeng Liu[1], Alexandra H. Brozena[1], Zongfu Yu[3], Jelena Srebric[2], Liangbing Hu[1]*

Passive daytime radiative cooling materials could reduce the energy needed for building cooling up to 60% by reflecting sunlight and emitting long-wave infrared (LWIR) radiation into the cold Universe (~3 kelvin). However, developing passive cooling structures that are both practical to manufacture and apply while also displaying long-term environmental stability is challenging. We developed a randomized photonic composite consisting of a microporous glass framework that features selective LWIR emission along with relatively high solar reflectance and aluminum oxide particles that strongly scatter sunlight and prevent densification of the porous structure during manufacturing. This microporous glass coating enables a temperature drop of ~3.5° and 4°C even under high-humidity conditions (up to 80%) during midday and nighttime, respectively. This radiative "cooling glass" coating maintains high solar reflectance even when exposed to harsh conditions, including water, ultraviolet radiation, soiling, and high temperatures.

Approximately 10% of annual global electricity use is devoted to the air-conditioning of buildings (*1*, *2*), and with cooling needs projected to triple by 2050, different approaches are required to mitigate strain on the electric grid and combat global warming (*3*). Passive daytime radiative cooling technologies can reduce these energy demands by applying materials to the building envelope that reflect >90% of solar radiation and emit heat through the atmospheric transparency window (8 to 13 μm) as long-wave infrared (LWIR) light into the cold Universe (~3 K) to achieve subambient temperatures (*4*–*7*). Beyond building use, passive radiative cooling technologies could also benefit solar cells (*8*), power plant condensers (*5*, *9*), high-performance textiles for personal thermal comfort (*10*, *11*), dew collection (*12*), and glacier melt mitigation (*13*, *14*). Various approaches have been demonstrated for passive radiative cooling based on nanophotonic structures that integrate multilayered inorganic thin films (e.g., composited ceramics and metals) (*15*–*18*). However, such structures require complex fabrication techniques with nanoscale precision (often in vacuum chambers), making them difficult and costly to scale, especially for building applications.

To improve the manufacturability of daytime radiative cooling materials, different approaches have been demonstrated with organic polymers, such as polymer-metal hybrid films (*19*, *20*), porous polymer coatings and foams (*10*, *21*, *22*), polymer-dielectric paints (*23*–*25*), delignified wood (*26*), and multilayer polymer films (*27*). However, organic polymers tend to yellow and degrade over time when exposed to ambient conditions (*17*, *28*), including ultraviolet (UV) light, heat, water, and ambient chemicals (e.g., NO_x, SO_x, and O_3), which increases the material's solar absorption, thus reducing the cooling power or even resulting in heating. In addition, the metals used in polymeric or inorganic radiative cooling structures, such as silver (*19*, *29*) and aluminum (*20*), can be oxidized or sulfated by chemical species found in air (O, O^-, Cl^-, H_2O_2, and SO_2) (*30*), which causes a reduction in solar reflectance (*31*, *32*). Clearly, environmental stability remains a roadblock in the practical implementation of this otherwise promising technology. And while environmentally stable micro- or nanoscale ceramics, such as SiO_2, SiC, and Al_2O_3, have been demonstrated for radiative cooling (*33*–*36*), their mechanical strength is poor owing to the lack of binders, making them unsuitable for buildings. Furthermore, to meet the needs of applications in more-extreme environments—for example, aerospace thermal control (*37*, *38*)—the passive radiative cooling materials need to withstand high temperatures of up to 1000°C without performance degradation.

In this study, we designed and demonstrated a solution-processed coating of photonic "cooling glass" that addresses the key challenges of polymer- and metal-based radiative cooling structures, being stable under ambient conditions, scalable, and low cost while demonstrating excellent passive cooling performance. We fabricated the cooling glass using a simple two-step process in which inexpensive glass and Al_2O_3 particles are mixed into a slurry that can be easily painted on a substrate, followed by thermal annealing of the low–melting point glass to realize the porous radiative cooling structure (Fig. 1A). The glass particles, which feature a low softening temperature (~350°C) and rich infrared-active vibrational modes in the atmospheric transparency window, serve as a nonconventional binder to form a robust porous supportive framework (characteristic size of ~12 μm; Fig. 1A) that can provide enhanced selective LWIR emission via Fröhlich resonances (*19*) w… scattering sunlight toward a fairly high s… reflectance (Fig. 1B). Additionally, Al_2O_3 particles (mean size of 0.5 μm) are mixed with the glass particles (Fig. 1A) to improve the composite's solar reflectance by allowing Mie scattering in the solar spectrum (Fig. 1B) while also serving as an anti-sintering agent to prevent the complete densification of the glass particles, which would otherwise result in a sunlight-transparent structure. This dual-particle design optimizes material and dimensional effects associated with passive radiative cooling, specifically combining a high solar reflectance of >0.96 and a high infrared emissivity of ~0.95 in the atmospheric transparency window. As a result, after applying this coating (~550 μm thick) to a 1.5-mm-thick glass substrate, outdoor experiments show that the substrate temperature can be 3.5°C lower than the ambient temperature (30°C) at noon (solar irradiance of 800 W/m^2) and 4°C lower at night (ambient temperature of 17.5°C), even in high humidity (up to 80%). Simulations of midrise apartment buildings indicate that coating the radiative cooling glass on roofing can enable an annual CO_2 emission reduction of ~10% given the decreased cooling needs. In addition to its excellent performance, the cooling glass features a high adhesion strength to substrates (e.g., tile, brick, glass, and metal) and maintains its high solar reflectance even when exposed to harsh conditions, including water, UV, soiling, and high-temperature flame shock (up to 1000°C). Therefore, this simple, scalable, cost-effective, and environmentally stable glass coating for passive radiative cooling can both reduce energy consumption to help mitigate global warming and improve residential comfort.

Modeling to guide the design

We selected low–melting point phosphate glass particles (a eutectic mixture of different dielectric oxides, including P_2O_5, SiO_2, Al_2O_3, Na_2O, K_2O, and B_2O_3) to form a framework for the cooling glass coating because of the material's high durability, low softening temperature (*39*), and rich infrared-active vibrational modes, such as Al–O, Si–O, P–O, etc. (fig. S1), which occur in the atmospheric transparency window. Additionally, to further improve the solar reflectance and achieve a porous glass structure, we chose α-Al_2O_3 particles featuring a high melting point (2072°C), high refractive index in the solar spectrum (~1.7, fig. S2), high bandgap (~7.0 eV) (*23*), and high thermodynamic and chemical stability, as well as low cost and abundance (*40*), to serve as an anti-sintering agent and thus ensure that the glass does not completely densify and become transparent upon heating. We first theoretically determined the optimal sizes for the

[1]Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA. [2]Department of Mechanical Engineering, University of Maryland, College Park, MD 20742, USA. [3]Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53705, USA.
*Corresponding author. Email: binghu@umd.edu
†These authors contributed equally to this work.

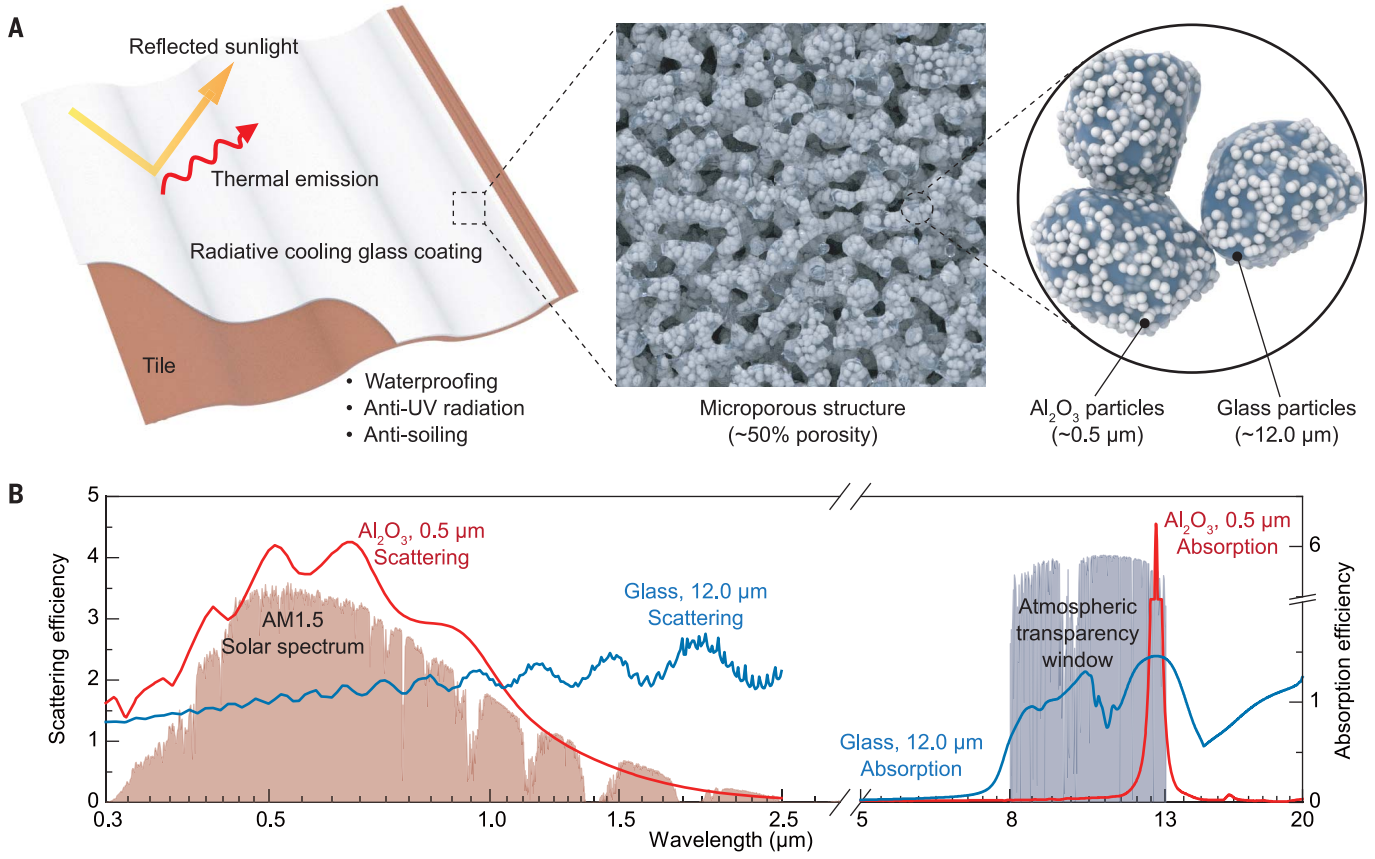


Fig. 1. An environmentally stable glass coating for daytime passive radiative cooling. (**A**) Schematic of the radiative cooling glass coating on a ceramic roofing tile, which can effectively reflect solar radiation (0.3 to 2.5 μm) and emit infrared radiation (i.e., thermal emission) to the cold sky through the atmospheric transparency window (8 to 13 μm). The radiative cooling glass coating features a porous structure (porosity: ~50%), in which low–melting point glass particles (mean diameter: ~6 μm; volume: ~30%) are partially sintered to form a framework decorated with Al_2O_3 particles (mean diameter: ~0.5 μm; volume: ~20%). The characteristic size of the glass clusters after sintering is ~12 μm. (**B**) Demonstration of the optical functionality of the glass particles and Al_2O_3 particles in the composite structure. The scattering and absorption efficiencies as a function of wavelength for glass (blue line) and Al_2O_3 (red line) particles were calculated on the basis of the Lorenz-Mie theory. The dual-particle design maximizes material and dimensional effects associated with passive radiative cooling, specifically high reflectance in the solar spectrum and high emissivity in the atmospheric transparency window.

glass and Al_2O_3 particles using Lorenz-Mie theory (*41*) to maximize the LWIR emittance and solar scattering efficiency, respectively (fig. S3). We calculated absorption efficiencies of individual spherical glass particles with various diameters as a function of wavelength (5 to 15 μm) (fig. S3A). The results indicate that a wide range of glass particle sizes (2 to 20 μm diameter) can lead to high absorption in the atmospheric transparency window, although a narrower diameter range of 8 to 15 μm maximizes the emission owing to phonon-enhanced Fröhlich resonances (figs. S3B and S4) (*19*). These glass particles also provide relatively high scattering efficiency across the solar spectrum (fig. S3C). We demonstrated the calculated scattering efficiencies of Al_2O_3 particles ranging in diameter from 0.3 to 3.0 μm (fig. S3D), suggesting that they scatter sunlight with wavelengths close to their diameters with the highest efficiency [i.e., Mie scattering (*11*)] (fig. S5). The Al_2O_3 particles with sizes of 0.5 to 0.7 μm exhibit the highest solar scattering efficiency across the entire solar spectrum (fig. S3E). Furthermore, these Al_2O_3 particles also exhibit a high absorption efficiency at wavelengths around 13 μm (fig. S3F), which indicates selective emissivity in the atmospheric transparency window. We further validated these predictions through full-wave simulations (figs. S6 to S8) using the Tidy3D software, which uses a graphics processing unit–accelerated finite-difference time-domain method (*42*).

Process and structure

To fabricate the porous photonic composite, we first dispersed glass particles with a mean diameter of 6.3 μm (2 to 15 μm range, softening temperature of 350°C; fig. S9) and Al_2O_3 particles with a mean diameter of 0.5 μm (0.3 to 1.0 μm range; fig. S10) at a 1:1 weight ratio in ethanol (fig. S11) and homogenously mixed by ball milling. We began with smaller-diameter glass particles as the starting material to account for the increase in size of the glass particles during sintering. The resulting slurry displayed good fluidity (Fig. 2A), with a viscosity that can be modified for different coating techniques, such as spray or brush coating (fig. S12), allowing the material to be applied on various substrates, such as brick (Fig. 2B), tile, metal, and glass, as well as scaled to multiple and large substrates (Fig. 2C). As a demonstration, we coated a ceramic tile with a mixture of the glass-Al_2O_3 slurry, allowed the solvent to evaporate, and then furnace-heated the sample at a temperature of ~600°C for ~1 min to selectively sinter the glass particles into a supportive framework (fig. S13). We used scanning electron microscopy (SEM) to observe the morphology of the glass and Al_2O_3 particle mixture before and after sintering. Before sintering, the glass and Al_2O_3 particles are homogeneously mixed, with the small-sized Al_2O_3 particles located at the interstices

and surfaces of the adjacent glass particles (Fig. 2D). During sintering, the reduced viscosity of the raw glass, ~100 Pa·s (*39*), facilitates rapid merging and bonding of the glass particles. This forms a framework of larger glass clusters (fig. S14), which adhere to the substrate. The Al_2O_3 particles are either affixed to the surface or enclosed within this structure (Fig. 2E and fig. S15). Notably, this dual-particle and partial-sintering process results in the formation of a microporous structure throughout the coating (Fig. 2, E and F). We further examined the porous structure of the glass-Al_2O_3 clusters by filling the voids with polymeric resin (*41*). As shown by cross-sectional SEM (Fig. 2G), the coating displays a porosity of ~50%. The clusters of glass and Al_2O_3 (light-gray regions) are connected and irregular, ranging in size from 5 to 20 μm, with an average size of ~12 μm (Fig. 2H), of which >60% were in the 8 to 15 μm range. Additionally, the pore sizes (Fig. 2H, corresponding to the darker-gray color shown in Fig. 2G) ranged from a few hundred nanometers to ~30 μm (average size of ~6.7 μm;

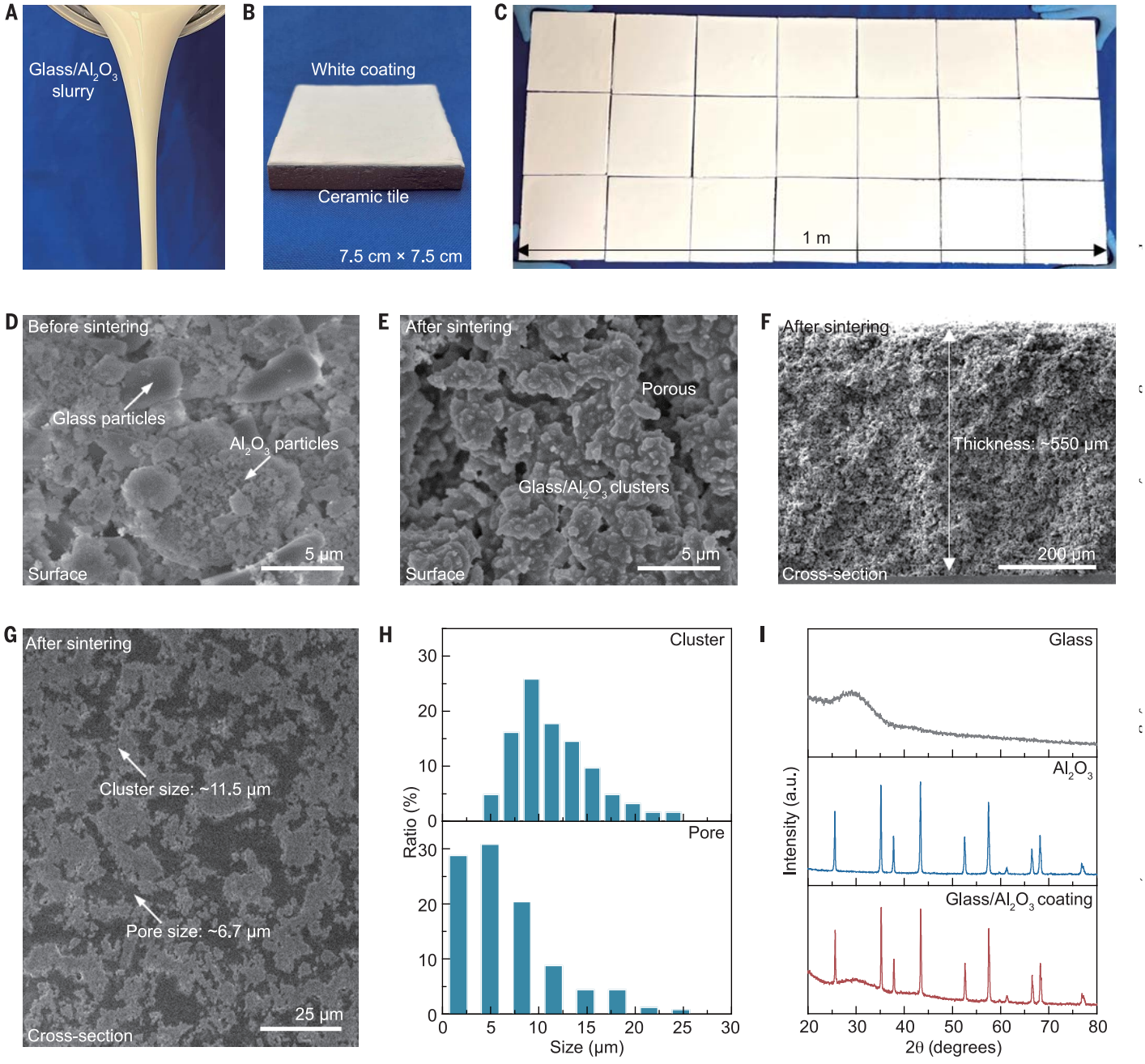


Fig. 2. Fabrication and morphology of the radiative cooling glass coating. (**A**) The glass-Al_2O_3 particle slurry in ethanol, which shows good fluidity. (**B**) The radiative cooling glass coating applied on a ceramic tile by means of sintering at ~600°C in a furnace. (**C**) The radiative cooling glass coating can be fabricated on a large scale. (**D** and **E**) SEM images of the glass and Al_2O_3 particle mixture (D) before and (E) after sintering at ~600°C for ~1 min. After the heat treatment, the Al_2O_3 particles are dispersed with the softened and merged glass particles, forming a cohesive microporous structure. (**F**) SEM image of the cross-sectional morphology of the microporous coating. (**G**) SEM image of the cross-sectional morphology of the microporous coating filled with polymeric resin (darker-gray color). (**H**) The size distribution of the glass-Al_2O_3 clusters and pores after sintering. Approximately 60% of the glass-Al_2O_3 clusters were in the 8 to 13 μm range, while ~80% of the pores were smaller than 10 μm. (**I**) XRD patterns of glass particles, Al_2O_3 particles, and sintered radiative cooling glass coating.

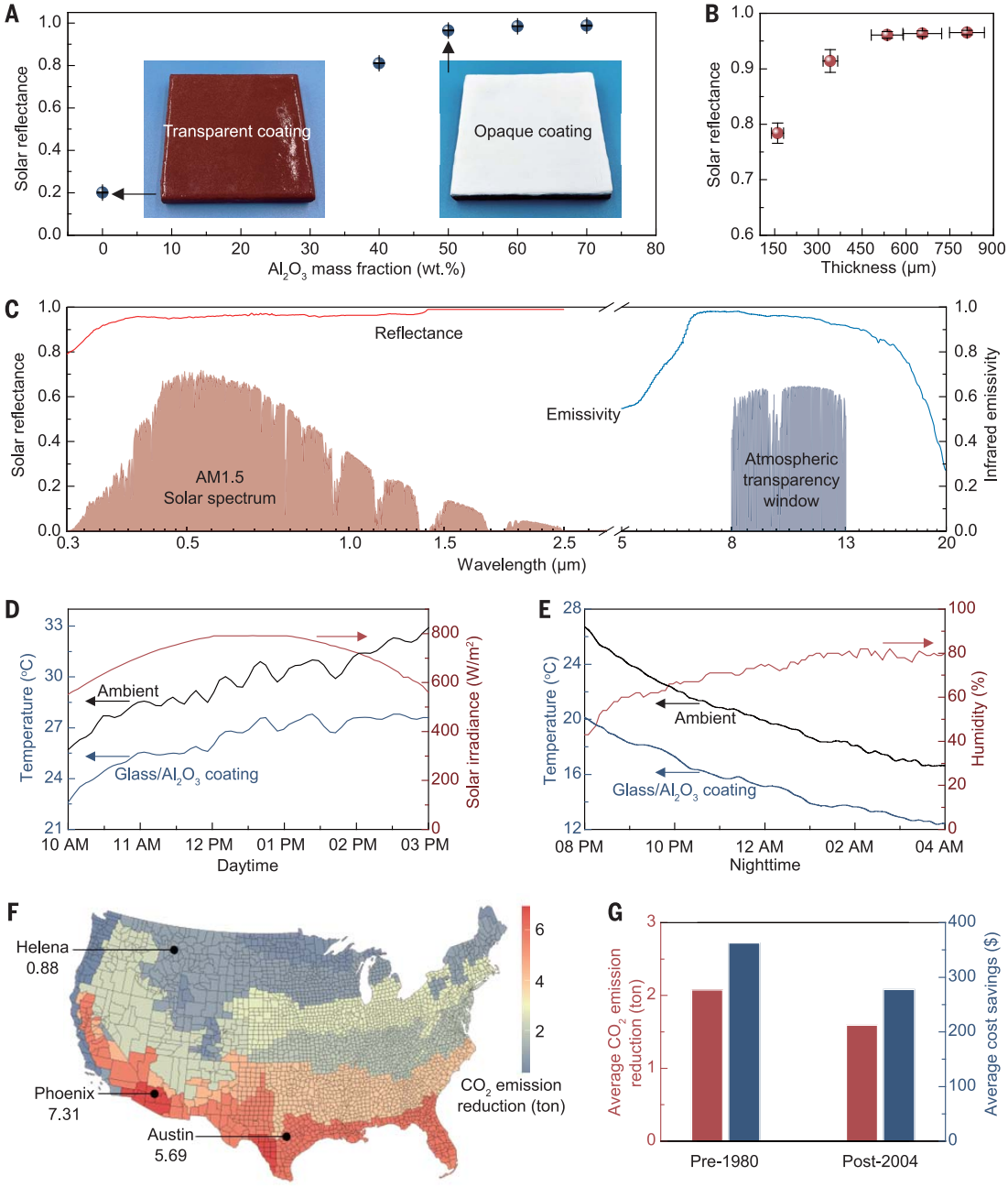


Fig. 3. The optical and thermal performance of the radiative cooling glass coating and simulated CO_2 emission reductions for roof applications. (**A**) Effect of the Al_2O_3 particle mass fraction on the solar reflectance of the radiative cooling glass coating with a thickness of ~550 μm. (**B**) Effect of the thickness of the radiative cooling glass coating on its solar reflectance, in which the mass fraction of the Al_2O_3 particles was ~50 wt %. (**C**) The measured solar reflectance and infrared emissivity of the radiative cooling glass coating with a thickness of ~550 μm in the range of 0.3 to 20 μm. (**D**) The measured temperature of a transparent glass substrate coated with the radiative cooling glass coating (~550 μm thick, 50 wt % Al_2O_3 and 50 wt % glass) under direct sunlight. The outdoor thermal performance was measured in College Park, Maryland, USA (38.9897°N, 76.9378°W; 22 m altitude) from 10 a.m. on 6 June 2022 to 4 a.m. on 7 June 2022. (**E**) Measured temperature of the transparent glass substrate covered by the radiative cooling glass coating at night for the same location. (**F**) Annual CO_2 emission reductions for cooling midrise apartment buildings (pre-1980) in US cities after applying the radiative cooling glass coating to the roof. (**G**) Average CO_2 reductions and cost savings for cooling the pre-1980 and post-2004 midrise apartment buildings after applying the radiative cooling glass coating to the roof.

fig. S16), with ~80% being smaller than 10 μm, which should also promote solar scattering and improve the solar reflectance by increasing the refractive index contrast (*43*). Note that x-ray diffraction (XRD) showed that the chemical structure of the Al_2O_3 and glass particles remained the same after sintering (Fig. 2I).

Optical performance optimization

We optimized the mass content of the Al_2O_3 particles in the cooling glass composite by examining the solar reflectance and infrared emissivity of the glass coating with different Al_2O_3 content (fig. S17). Without Al_2O_3 particles, the glass particles merge during the heat treatment, eliminating most of the pores and forming a transparent coating with a very low solar reflectance of ~0.2 (Fig. 3A, inset). With the addition of Al_2O_3 particles (e.g., 40 wt %; fig. S17B), the coating becomes slightly porous (fig. S18), and its solar reflectance improves remarkably (to ~0.8; Fig. 3A). When the mass fraction of the Al_2O_3 particles increases to 50 wt %, the resulting coating (thickness of ~550 μm) forms a porous structure (~50% porosity; Fig. 2G) and demonstrates a high solar reflectance of ~0.96 (Fig. 3A) as a result of the Al_2O_3 preventing the glass particles from completely merging. When we increased the mass fraction of the Al_2O_3 to >60 wt %, the solar reflectance of the coating reached ~0.98 (Fig. 3A). However, the resulting coating showed low structural strength and resistance to abrasion and impact because of the lower content of the glass binder, weakening the interconnections of the supportive glass backbone. These results suggest that a 40 to 60 wt % Al_2O_3 content

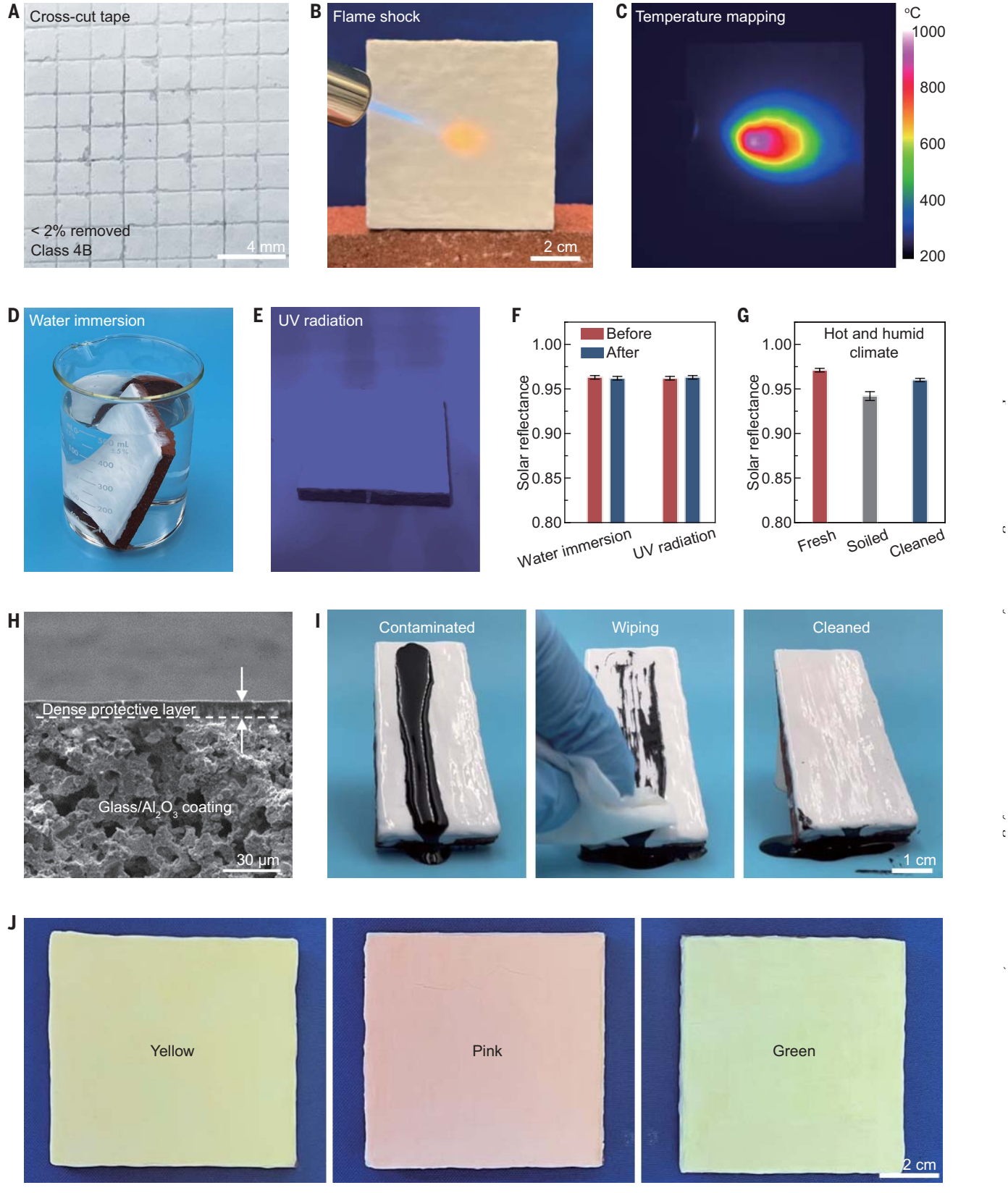


Fig. 4. Environmental stability of the radiative cooling glass. (**A**) Photograph of the radiative cooling glass after a standard cross-cut tape test (ASTM D3359-2, Method A; film thickness: ~100 μm). (**B** and **C**) The cooling glass coating exposed to flame shock (~10 s) with a temperature of up to 1000°C. (**D** and **E**) Photographs of the radiative cooling glass coating immersed in (D) water and exposed to (E) UV radiation with an intensity of ~5.0 W/cm^2. (**F**) Comparison of

the solar reflectance of the radiative cooling glass coating after immersion in water for 60 days and UV radiation for 80 days. (**G**) Comparison of the solar reflectance of the radiative cooling glass coatings before and after laboratory soiling tests (ASTM D7897-18) simulating a hot and humid climate (e.g., Miami, Florida). (**H**) SEM image of the radiative cooling glass coating covered by a dense and transparent protective layer of thin solid glass with a thickness of ~8 μm. (**I**) Demonstration of the anti-contamination ability of the radiative glass coating with a transparent protective layer. Owing to the dense transparent protective layer, liquid pollutant on the radiative cooling glass can be easily wiped off, showing excellent anti-contamination functionality. (**J**) Photographs of ceramic tiles coated with various colored radiative cooling glass coatings.

is an optimal range to ensure a coating with good stability and high solar reflectance. Meanwhile, we found that the infrared emissivity of the coating (thickness of ~550 μm) varies very little with the mass fraction of Al_2O_3 in the range of 40 to 70 wt % (fig. S19). We also studied the influence of the glass coating thickness and found that the solar reflectance was ~0.78 when the thickness was ~150 μm and reached 0.961 when we increased the thickness to ~550 μm (Fig. 3B). These results indicate that the thickness of the coating must be >500 μm (fig. S20) to achieve a solar reflectance of >0.95 for high subambient radiative cooling performance under a solar irradiance of ~1000 W/m^2 (*4*). For coatings of 500 μm or greater, the solar reflectance was the same, regardless of whether the substrate material was opaque or transparent (figs. S21 and S22). The coating methods also have an influence on the solar reflectance of the glass coatings. Compared with brush-coated glass coatings, spray-coated glass coatings (fig. S23) have a rougher, more porous structure, which results in slightly lower solar reflectance; however, increasing the Al_2O_3 mass fraction to ~60 wt % in spray-coated coatings can achieve solar reflectance of >0.96 with thickness of around 500 μm. We measured the reflectance and emissivity spectra of the optimized cooling glass coating (50 wt % Al_2O_3 and 50 wt % glass, thickness of ~550 μm) in the solar (0.3 to 2.5 μm) and infrared (5 to 20 μm) regions, respectively (Fig. 3C). The radiative cooling glass coating (Fig. 3C, red line) exhibits a high solar reflectance of 0.96 and high emissivity (~0.95, blue line) in the atmospheric transparency window (8 to 13 μm), allowing it to minimize solar heating while simultaneously radiating heat to the cold sink of outer space in the form of LWIR radiation. Additionally, we demonstrated that our process is compatible with a wide range of dielectric particles, such as TiO_2, ZnO, and BN (fig. S24). Moreover, the process can be effectively applied to curved surfaces (fig. S25) and can be easily scaled up (fig. S26) for large-scale applications (fig. S27).

Radiative cooling performance

We experimentally measured (fig. S28) the subambient radiative cooling performance of the cooling glass coating coated on a transparent glass substrate with a thickness of around 1.5 mm during the day (Fig. 3D) and night (Fig. 3E) in College Park, Maryland, USA (38.9897°N, 76.9378°W; 22 m altitude), which has a humid subtropical climate (high humidity, climate zone 4A) (*44*). Figure 3D shows that at midday (between 12 and 1 p.m.), the temperature of the glass coating was 3.5°C lower than the ambient temperature of 30°C at 30% relative humidity and around 1.5 m/s wind speed (fig. S29) and a solar irradiance of ~790 W/m^2. At night (between 2 and 4 a.m.; Fig. 3E), the glass coating was ~4°C cooler than the surrounding temperature of 17.5°C at a high humidity of up to 80% and a wind speed of 0.7 m/s (fig. S30). We used a theoretical model (*39*, *41*) (fig. S31) to estimate that the cooling power of the glass coating is ~60 W/m^2 (fig. S32). The thermal performance of the glass coating is comparable to that of radiative cooling materials with similar optical properties (solar reflectance of >0.96, emissivity of ~0.95) that were tested in humid locations (*29*, *45*). These outdoor experiments confirmed that the developed glass coating shows high radiative cooling performance during both day and night, even in high-humidity environments.

CO_2 emission reduction for cooling

To evaluate the potential reduction in CO_2 emissions for cooling and to determine the most effective use of the cooling glass coating, we used EnergyPlus (*46*, *47*), a well-validated whole-building energy simulator, to calculate the equivalent CO_2 emission reductions and utility cost savings that could be achieved by applying the glass coating to the roof of a building. We investigated its performance in all continental US cities, with a focus on 16 cities representing all climate zones, including Albuquerque (New Mexico), Atlanta (Georgia), Austin (Texas), Boulder (Colorado), Chicago (Illinois), Duluth (Minnesota), Fairbanks (Alaska), Helena (Montana), Honolulu (Hawai'i), Las Vegas (Nevada), Los Angeles (California), Minneapolis (Minnesota), New York City (New York), Phoenix (Arizona), San Francisco (California), and Seattle (Washington). We conducted simulations using a reference midrise apartment building (roof area of 783 m^2, floor area of 3135 m^2, and window-to-wall ratio of 20%) provided by the US Department of Energy (fig. S33). Both old (pre-1980) and new (post-2004) structures were studied. To determine the relevant CO_2 emission reductions and cost savings in different cities, we treated this reference apartment building as the baseline (fig. S34), then modified the optical properties (Fig. 3C) of the roof on the basis of our measurements (fig. S35). Figure 3F and fig. S36 show the color maps of the annual CO_2 emission reductions for cooling the pre-1980 and post-2004 reference midrise apartment buildings in US cities when the radiative cooling glass coating is applied to the roof, respectively. More-detailed data on the annual energy savings, CO_2 emission reductions, and cost savings for the representative 16 cities can be found in fig. S37. Cities (Fig. 3F) that experience hot summers had the highest CO_2 emission reductions—Phoenix (7.31 tons) and Austin (5.69 tons), for example—particularly compared with cities that have cold winters, such as Helena (0.88 tons). Pre-1980 buildings had greater CO_2 reductions than post-2004 buildings because the walls are less well insulated, with an average reduction of 2.1 tons (~8.3%) for old buildings and 1.6 tons (~7.6%) for new buildings compared with the baseline (Fig. 3G). Applying the cooling glass coating on the roof of the midrise apartment building also resulted in an average annual cost savings of ~$350 for old buildings and ~$290 for new buildings compared with the baseline (Fig. 3G).

Environmental stability and color

We also evaluated the environmental stability of the cooling glass to determine its long-term viability. We first tested the adhesion strength of the glass coating on ceramic tile using a standard cross-cut tape test [ASTM D3359 (*48*)] (Fig. 4A). After the test, the area of the glass coating that was removed was <2%, demonstrating a high Class 4B adhesion strength (*48*). We also found that the cooling glass is flame resistant and can withstand high temperature flame shock (for ~10 s) of up to 1000°C (Fig. 4, B and C), with no resulting change in the material's microstructure or optical performance reflectance (fig. S38). In general, although the emissivity of radiative cooling materials tends to be stable, the solar reflectance is sensitive to ambient environments, decaying over time and reducing the passive cooling capability (*28*). However, we did not see obvious changes in the solar reflectance of the cooling glass after exposure to water for 60 days (Fig. 4D), followed by UV irradiation [~5.0 W/cm^2 for 80 days, equivalent to 3 years of outdoor radiation in the Florida area (*49*, *50*)] (Fig. 4E), which we attribute to the coating's environmentally stable inorganic

materials (Fig. 4F and fig. S39). Additionally, we conducted a dynamic test to simulate raindrop impact using jet impingement for 24 hours (*41*). The results revealed that our radiative cooling glass coating exhibited a <1% change in both solar reflectance and mass, indicating its resilience to raindrops (fig. S40). We also performed soiling tests (*41*) that simulated conditions in a hot and humid climate (e.g., Miami, Florida), following ASTM D7897-18 (*51*), as shown in fig. S41. A 3% decrease in solar reflectance was observed after soiling (from 97% down to 94%) (Fig. 4G and fig. S42). However, the solar reflectance returned to 96% after the coating was cleaned with forced air flow (10 m/s). We attribute the 1% reduction in solar reflectance to the simulated soiling material becoming trapped in the coating's porous structure. For hot and dry (e.g., Phoenix, Arizona) and moderate climate (e.g., Youngstown, Ohio) soiling conditions, we also observed a ~1% reduction in solar reflectance after soiling (figs. S43 and S44). In addition to laboratory tests, our outdoor experiments indicate that the solar reflectance of the coating remained nearly unchanged after two months of exposure, even in the presence of visible black particles on the tile surfaces resulting from wildfires (fig. S45).

To further improve the soiling resistance and environmental stability of the glass coating, we added another layer (<10 μm thick; Fig. 4H and fig. S46) of the low–melting point glass particles on top of the porous cooling glass coating to produce a thin and dense transparent glass surface after heating (*41*). The coating appears shiny because of the specular reflection of the dense transparent protective glass layer (fig. S46A). The added transparent protective layer reduces the solar reflectance from ~0.963 to ~0.95 (fig. S46E), because of interference between light reflections from different interfaces (i.e., air/dense glass/porous glass coating) (figs. S47 and S48). However, this reflectance is still higher or comparable to that of previously reported daytime radiative cooling structures (*11*, *33*, *35*). Additionally, the dense structure of the transparent protective layer allows liquid pollutant (high-concentration carbon nanoparticles dispersed in water) to be easily cleaned from the cooling glass coating, providing excellent anti-contamination functionality (Fig. 4I). In contrast, when we applied the same liquid pollutant to the unprotected cooling glass, the contaminant is able to permeate the porous structure, markedly reducing the solar reflectance (fig. S49). These results suggest that a thin, transparent layer of glass on the surface of the porous cooling glass is sufficient to remarkably improve the material's long-term stability against environmental exposure. Furthermore, we showed that when combined with a transparent protective layer, the radiative cooling glass coating displays enhanced hardness (fig. S50).

Aesthetic appeal is one of the aspects that could limit the widespread application of radiative cooling materials, which are generally limited to the color white. We showed that the cosmetic appearance of the radiative cooling glass coating can be modified by simply mixing the glass and Al_2O_3 precursor particles with inorganic dyes such as phosphors (fig. S51). Figure 4J shows colored radiative cooling glass coatings where 10 wt % phosphors, including nitride($CaAlSiN_3:Eu^{2+}$), $(Sr,Ba)_2SiO_4:Eu^{2+}$, and (Ce,Gd):YAG, were adopted to produce pink, green, and yellow colors, respectively. All the colored glass coatings have a solar reflectance of >0.9, and the value of the yellow coating can reach up to 0.95 (fig. S51).

Conclusions

Here we report a dual-particle design approach to develop environmentally stable and cost-effective microporous photonic glass coatings for daytime radiative cooling that avoids the use of organic polymers, metals, and complex multilayered structures. The design features large glass particles (~12 μm in diameter) with a size well within the atmospheric transparency window to form a porous framework that provides enhanced selective LWIR emission while scattering sunlight toward a high solar reflectance. Meanwhile, smaller Al_2O_3 particles (~0.5 μm in diameter) are incorporated to improve the composite's solar reflectance and prevent the complete densification of the glass particles. The radiative cooling glass coating is compatible with a number of different combinations of glass framework featuring different softening points (300° to 1000°C) and dielectric particles (e.g., TiO_2, ZnO, BN, and Al_2O_3) and can be easily extended to develop colored radiative cooling glass coatings by incorporating dyes. This solution-based process is scalable and can be easily applied to various surfaces, including roofs and walls, by means of brush or spray coating, while the sintering treatment can be achieved through a variety of means (e.g., hot air gun, heat lamp). Note that the mild sintering is to simply soften the glass sufficiently to generate a cohesive structure of optimized cluster size, in part dictated by the size of the particles. The excellent optical properties (i.e., >0.96 solar reflectance and ~0.95 selective LWIR emissivity) and environmental stability enable the glass coating to provide outstanding long-term radiative cooling performance even in high-humidity environments. This work provides a radiative cooling structure with high solar reflectance, high selective LWIR, high resistance to ambient degradation, and high operating temperature that is readily available for large-scale and long-term deployment, such as on buildings, data centers, and for cold-chain transportation, as well as applications in more-extreme environments (e.g., aerospace).

REFERENCES AND NOTES

1. United Nations Environment Programme (UNEP) and International Energy Agency (IEA), *Cooling Emissions and Policy Synthesis Report: Benefits of cooling efficiency and the Kigali Amendment* (UNEP and IEA, 2020).
2. J. Woods *et al*., *Joule* **6**, 726–741 (2022).
3. International Energy Agency (IEA), *The Future of Cooling: Opportunities for Energy Efficient Air Conditioning* (IEA, 2018); https://doi.org/10.1787/9789264301993-en.
4. X. Yin, R. Yang, G. Tan, S. Fan, *Science* **370**, 786–791 (2020).
5. D. Zhao *et al*., *Appl. Phys. Rev.* **6**, 021306 (2019).
6. E. Teitelbaum *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 21162–21169 (2020).
7. J. Kośny, D. W. Yarbrough, in *Thermal Insulation and Radiation Control Technologies for Buildings*, J. Kośny, D. W. Yarbrough, Eds. (Springer, 2022), pp. 1–35.
8. L. Zhu, A. Raman, K. X. Wang, M. A. Anoma, S. Fan, *Optica* **1**, 32–38 (2014).
9. A. Aili, D. Zhao, G. Tan, X. Yin, R. Yang, *Appl. Energy* **302**, 117515 (2021).
10. P.-C. Hsu *et al*., *Science* **353**, 1019–1023 (2016).
11. S. Zeng *et al*., *Science* **373**, 692–696 (2021).
12. M. Zhou *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2019292118 (2021).
13. J. N. Munday, *Joule* **3**, 2057–2060 (2019).
14. J. Li *et al*., *Sci. Adv.* **8**, eabj9756 (2022).
15. A. P. Raman, M. A. Anoma, L. Zhu, E. Rephaeli, S. Fan, *Nature* **515**, 540–544 (2014).
16. M. M. Hossain, B. Jia, M. Gu, *Adv. Opt. Mater.* **3**, 1047–1051 (2015).
17. D. Chae *et al*., *ACS Appl. Mater. Interfaces* **12**, 8073–8081 (2020).
18. K. Tang *et al*., *Science* **374**, 1504–1509 (2021).
19. Y. Zhai *et al*., *Science* **355**, 1062–1066 (2017).
20. L. Zhou *et al*., *Nat. Sustain.* **2**, 718–724 (2019).
21. J. Mandal *et al*., *Science* **362**, 315–319 (2018).
22. A. Leroy *et al*., *Sci. Adv.* **5**, eaat9480 (2019).
23. J. Mandal, Y. Yang, N. Yu, A. P. Raman, *Joule* **4**, 1350–1356 (2020).
24. X. Li *et al*., *Cell Rep. Phys. Sci.* **1**, 100221 (2020).
25. X. Li, J. Peoples, P. Yao, X. Ruan, *ACS Appl. Mater. Interfaces* **13**, 21733–21739 (2021).
26. T. Li *et al*., *Science* **364**, 760–763 (2019).
27. S. Wang *et al*., *Science* **374**, 1501–1504 (2021).
28. H. Takebayashi *et al*., *Energy Build.* **114**, 173–179 (2016).
29. C. Lin *et al*., *Adv. Mater.* **34**, e2109350 (2022).
30. H.-Y. Wu *et al*., *Nanomaterials* **12**, 1054 (2022).
31. A. García-Segura, A. Fernández-García, M. Ariza, F. Sutter, L. Valenzuela, *Renew. Sustain. Energy Rev.* **62**, 453–467 (2016).
32. R. Bethea, M. Barriger, P. Williams, S. Chin, *Sol. Energy* **27**, 497–511 (1981).
33. H. Bao *et al*., *Sol. Energy Mater. Sol. Cells* **168**, 78–84 (2017).
34. A. R. Gentle, G. B. Smith, *Nano Lett.* **10**, 373–379 (2010).
35. S. Atiganyanun *et al*., *ACS Photonics* **5**, 1181–1187 (2018).
36. T. Li *et al*., *Chem. Eng. J.* **452**, 139518 (2023).
37. T. G. Hawarden *et al*., in *Infrared and Submillimeter Space Missions in the Coming Decade*, H. A. Thronson, M. Sauvage, P. Gallais, L. Vigroux, Eds. (Springer Dordrecht, 1995), pp. 45–56.
38. M. Holynska *et al*., *J. Spacecr. Rockets* **56**, 1358–1370 (2019).
39. H. Masai *et al*., *Sci. Rep.* **11**, 214 (2021).
40. W. Jin, P. K. Chu, in *Encyclopedia of Biomedical Engineering: Volume 2*, R. Narayan, Ed. (Elsevier, 2019), pp. 425–439.
41. See supplementary materials.
42. Flexcompute, Tidy3D: Next-gen Electromagnetic Simulation Tool, Example Library: Nanophotonics—Radiative cooling glass coating; https://www.flexcompute.com/tidy3d/examples/notebooks/RadiativeCoolingGlass/.

43. M. P. Diebold, *Application of Light Scattering to Coatings: A User's Guide* (Springer, 2014).
44. M. C. Baechler *et al*., Tech. Rep. PNNL-17211 (Pacific Northwest National Laboratory, 2010); https://doi.org/10.2172/1068658.
45. H. Zhong *et al*., *ACS Appl. Mater. Interfaces* **12**, 51409–51417 (2020).
46. EnergyPlus software; https://energyplus.net/.
47. P. C. Tabares-Velasco, C. Christensen, M. Bianchi, *Build. Environ.* **54**, 186–196 (2012).
48. ASTM Standard D3359-22, "Standard Test Methods for Rating Adhesion by Tape Test" (ASTM International, 2023); https://doi.org/10.1520/D3359-22.
49. M. M. Shokrieh, A. Bayat, *J. Compos. Mater.* **41**, 2443–2455 (2007).
50. L. Jacques, *Prog. Polym. Sci.* **25**, 1337–1362 (2000).
51. ASTM Standard D7897-18, "Standard Practice for Laboratory Soiling and Weathering of Roofing Materials to Simulate Effects of Natural Exposure on Solar Reflectance and Thermal Emittance" (ASTM International, 2018); https://doi.org/10.1520/D7897-18.

ACKNOWLEDGMENTS

L.H. acknowledges support from the University of Maryland A. James Clark School of Engineering. We acknowledge the use and support of the Maryland NanoCenter and its AIMLab. **Funding:** This project was not directly funded. **Author contributions:** L.H. and X.Z. conceived the idea and designed the present work. All authors contributed to experiments and data analysis. X.Z., L.H., and A.H.B. wrote the manuscript. **Competing interests:** L.H. and X.Z. report a pending patent application, filed on 11 November 2022 (US Provisional Application No. 63/383328). Z.Y. is a cofounder of Flexcompute Inc., which develops the photonic simulation software used in this paper. The remaining authors declare no competing interests. **Data and materials availability:** The data that support the findings of this study are available within this article and its supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi2224
Materials and Methods
Supplementary Text
Figs. S1 to S51
Tables S1 to S4
References (*52*–*73*)

Submitted 11 April 2023; accepted 19 September 2023
10.1126/science.adi2224

RADIATIVE COOLING

Hierarchically structured passive radiative cooling ceramic with high solar reflectivity

Kaixin Lin[1], Siru Chen[1], Yijun Zeng[2,3], Tsz Chung Ho[1], Yihao Zhu[1], Xiong Wang[2], Fayu Liu[2,3], Baoling Huang[4], Christopher Yu-Hang Chao[3,5], Zuankai Wang[3]*, Chi Yan Tso[1]*

Passive radiative cooling using nanophotonic structures is limited by its high cost and poor compatibility with existing end uses, whereas polymeric photonic alternatives lack weather resistance and effective solar reflection. We developed a cellular ceramic that can achieve highly efficient light scattering and a near-perfect solar reflectivity of 99.6%. These qualities, coupled with high thermal emissivity, allow the ceramic to provide continuous subambient cooling in an outdoor setting with a cooling power of >130 watts per square meter at noon, demonstrating energy-saving potential on a worldwide scale. The color, weather resistance, mechanical robustness, and ability to depress the Leidenfrost effect are key features ensuring the durable and versatile nature of the cooling ceramic, thereby facilitating its commercialization in various applications, particularly building construction.

Energy used for cooling continues to rise. Carbon dioxide emissions resulting from space cooling have more than doubled to almost 1 billion metric tons over the past 30 years (*1*). Unfortunately, the carbon footprint of cooling systems contributes to global warming, creating a vicious cycle further exacerbating the need for air-conditioning. To mitigate these environmental concerns, renewable energy sources, such as wind, tidal, and solar power, have been harnessed to meet the ever-increasing energy demand. However, the development and implementation of these energy systems require extensive land use and costly installations. In contrast, passive radiative cooling (PRC), in which surfaces use the cryogenic Universe as a natural heat sink for heat removal, provides an energy-neutral solution for energy transfer and space cooling (*2*–*4*).

In recent decades, radiative cooling systems (coolers) of various photonic architectures with subwavelength-scale dimensions have been designed for tuning the thermal spectrum (*5*–*17*). Although the tailored long-wave infrared spectrum helps to suppress heat gain from atmospheric radiation, the sophisticated material composition for thermal selectivity sacrifices solar reflectivity (R_{solar}). In addition, even though these designs can easily achieve deep nighttime subambient cooling, they struggle to generate cooling during periods of high cooling demand in the daytime.

Atmospheric window (AW) transmittance (τ_{AW}), which serves as the only avenue for radiative heat transfer between the cooler and the Universe, is highly affected by the local climate and environment (*18*–*20*). On the other hand, solar radiation dominates the radiative heat exchange of a cooler during daytime given that the Sun (5800 K) emits electromagnetic radiation at an intensity that is 11 orders of magnitude higher than that of a room temperature object (288 K). We numerically solved the steady-state heat transfer model for a sky-facing cooler, enabling estimation of the impact of a cooler's optical properties, that is, R_{solar} and AW emissivity (ε_{AW}), on the cooling performance (fig. S1). The model assumes a constant surrounding air temperature and solar intensity, so that the cooler's cooling power changes according to the cooler's optical properties. The calculation results (fig. S1) clearly indicate that enhancing R_{solar} can yield a higher cooling capacity than can enhancing ε_{AW} by the same amount, regardless of τ_{AW}.

However, the development of a daytime PRC material with both high R_{solar} and practicality for outdoor application is challenging. In contrast to patterned and periodic optical designs that are small-scale and costly, random media, such as particles (*21*–*27*) and micropores (*28*–*35*), have recently gained attention for achieving desirable daytime PRC. Researchers have investigated disordered designs, such as those of particle-doped polymers (*21*–*27*), aerogels (*28*), porous membranes (*29*–*33*), fibers (*34*, *36*), and delignified wood (*35*). An inherent advantage of these designs is their ability to obviate the need for intricate parameter control, affording simplicity and scalability. However, solar absorption by these materials is inevitable. The dielectric pigments (e.g., titanium dioxide and zinc oxide) and polymers used in these disordered designs absorb sunlight within different wavelength ranges (*37*, *38*). In addition, a high content of organic material potentially leads to degradation during long-term outdoor use (*39*). For potential application scenarios, such as buildings, this could lead to frequent replacement or maintenance of the building envelope and result in exorbitant cost. The trade-off between the optical properties and applicability has always posed a challenge for practical implementation of PRC technology in the real world. Recently, PRC designs using pure inorganic materials to achieve reliable stability offer insight for addressing this challenge (*40*–*42*). In this work, we designed a hierarchically porous PRC material in the form of a bulk ceramic, which can be easily obtained using accessible inorganic materials through facile fabrication processes. The design simultaneously achieves near-ideal R_{solar} and robust applicability, showing great potential for application in the real world.

Bio-whiteness

Cyphochilus, a beetle native to Southeast Asia, is the whitest known insect on Earth. We learned from the intricate biological structure of this beetle how to design a robust ceramic-format cooler. On the basis of the investigation of the scattering system of the beetle's scales, our cooling ceramic was engineered with a hierarchically porous structure that led to a near-ideal R_{solar}. This cooler is easily fabricated, requiring neither precision instruments nor meticulous regulation of parameters, and it boasts excellent daytime cooling performance, thus reducing energy consumption for indoor cooling. In the vast forest ecosystem, the beetle's white appearance serves as camouflage against fungal backgrounds, shielding it from predators. The *Cyphochilus* specimen used for morphological and optical characterization (Fig. 1A) is distinguished by a bright white exoskeleton comprising the head, thorax, and forewings. Upon closer examination using scanning electron microscopy (SEM), we detected the very small scales that densely cover the white areas, with ~15,000 scales/cm^2 (Fig. 1B). These scales, shaped like teardrops, are only 6 μm thick and comprise a highly connected random network of chitin with a filling fraction of ~60% and a typical diameter of 0.25 ± 0.05 μm (Fig. 1C and fig. S2). Spectroscopic characterization revealed that the scale-covered skeleton exhibits an average R_{solar} of >60%, in stark contrast to the R_{solar} value of only 32% for a scale-free skeleton (Fig. 1D and fig. S3). Notably, the scale covering contributed more than 40% to the reflectivity within the visible (VIS) range (0.38 to 0.75 μm), which is responsible for the beetle's brilliant whiteness. The exceptional visual whiteness of such a thin structure suggests the potential for the development of efficient R_{solar} in daytime PRC applications.

In the plastic, ink, and paint industries, products are often rendered white by incorporating pigments with high refractive index values into polymers. Similarly, the whiteness of *Cyphochilus* beetle scales can be attributed to a scattering system comprising a chitin filament network

[1]School of Energy and Environment, City University of Hong Kong, Hong Kong, China. [2]Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China. [3]Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China. [4]Department of Mechanical and Aerospace Engineering, Hong Kong University of Science and Technology, Hong Kong, China. [5]Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China.
*Corresponding author. Email: chiytso@cityu.edu.hk (C.Y.T.); zk.wang@polyu.edu.hk (Z.W.)

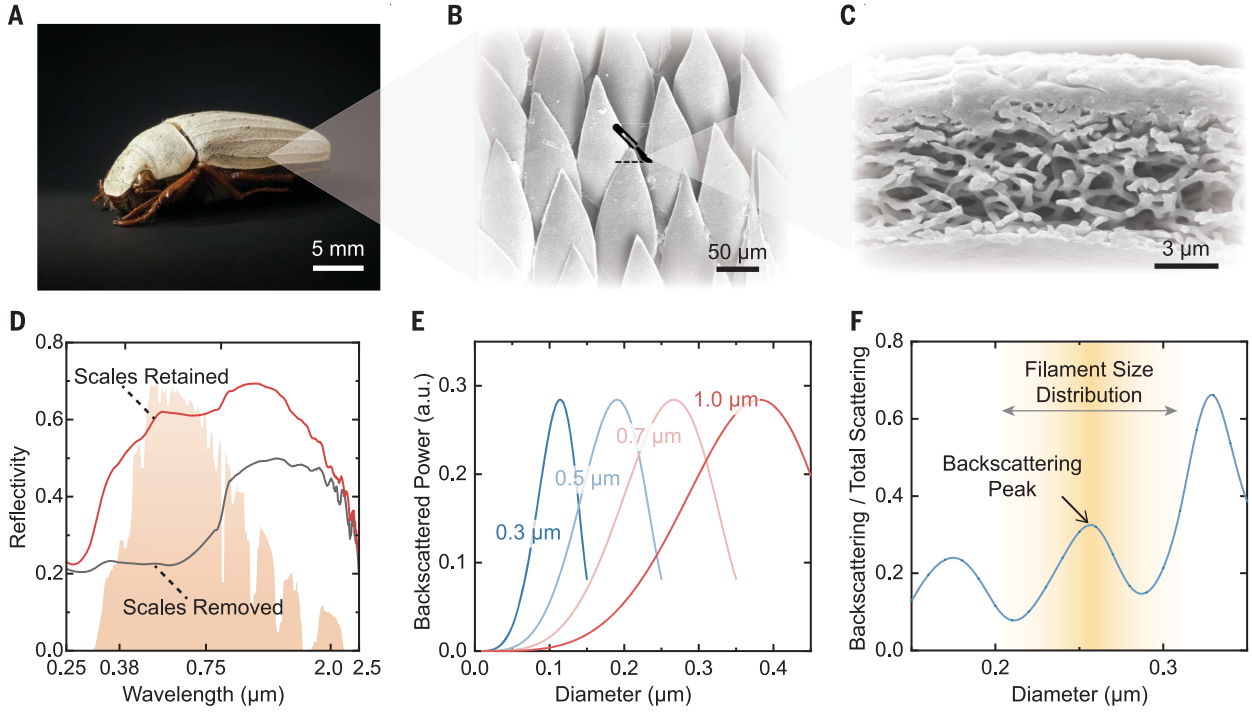


Fig. 1. Whitest beetle in the world. (**A**) Photograph of the *Cyphochilus* specimen demonstrating its white appearance. (**B**) SEM image of *Cyphochilus* scales. (**C**) Cross-sectional SEM image of a *Cyphochilus* scale. (**D**) Measured R_{solar} spectra of *Cyphochilus* exoskeleton surfaces with and without scale covering. (**E**) Simulated backscattered power for incident light of four wavelengths (0.3, 0.5, 0.7, and 1.0 μm) as a function of the chitin filament diameter. (**F**) Simulated scattering center dimension–dependent variation in the ratio of the backscattered power and the total scattered power for 0.5 μm incident light. The orange shading indicates the size distribution of the chitin filament.

and a complementary interconnected structure of air pockets. To better understand the scale whiteness, we modeled the Mie scattering behavior of chitin filament using Maxwell's equations (fig. S4). Considering a refractive index value of 1.56 to simulate chitin (*43*), we calculated the backscattering intensity at four incident wavelengths covering the ultraviolet (UV) to near-infrared (NIR) ranges (Fig. 1E). The results show that chitin filament with a diameter of 0.3 μm strongly scatters light with blue wavelengths, and as the size of chitin filament increases, the peak of the scattering intensity shifts toward red wavelengths (Fig. 1E). Because most *Cyphochilus* chitin filaments are between 0.2 and 0.3 μm in size, the scale most efficiently scatters light between 0.5 and 0.7 μm, which results in high opaqueness within the visible range. Additionally, we calculated the ratio of the backscattering intensity to the overall scattering for 500-nm incident light and found that the beetle's chitin filament dimension is notably coupled with one of the backscattering peaks located at 0.25 μm, confirming that the size of chitin filament is optimized for maximum scattering of visible light to achieve high visual whiteness (Fig. 1F).

High–solar reflectivity cooling ceramic

On the basis of the biostructure analysis, we engineered alumina particles through phase inversion and sintering to obtain the cooling ceramic (Fig. 2A and fig. S5). A three-component homogeneous solution, composed of polyethersulfone (PES), *N*-methyl-2-pyrrolidone (NMP), and α-alumina, was cast on a flat substrate and immersed in ethanol. The phase inversion takes place when the ethanol diffuses into the casting and dissolves with NMP. This process results in a polymer-rich membrane, forming an anisotropic porous network carrying alumina particles. Note that the porous alumina-polymer membrane can be flexible and imprinted with patterns or shaped into any form desired. In a high-temperature sintering process, PES starts to combust above 500°C. Both energy-dispersive spectroscopy (EDS) mapping (fig. S6A) and thermogravimetric analysis (TGA) (fig. S6B) confirmed the complete removal of PES through the sintering process. Meanwhile, alumina particles were bonded, resulting in a cooling ceramic with a well-preserved porous structure. Manipulating the alumina concentration in the phase inversion precursor solution yielded a short transport mean free path leading to high R_{solar} of the thin cooling ceramic (fig. S7). The as-obtained cooling ceramic shows intense whiteness under natural light (Fig. 2A) and has a structure similar to that of *Cyphochilus* scales, with a densely packed outer layer and numerous internal voids, but with a wider structural dimension distribution (Fig. 2, B and C). The cooling ceramic exhibits near-perfect reflectivity in the UV, VIS, and NIR ranges, leading to an R_{solar} value of 99.6%, versus a value of 89.5% for silver, a value of 88.6% for white-pigmented polymer, and a value of 76.2% for white commercial tiles (Fig. 2D). For scattering systems containing a certain volume of pores, a multidispersed pore system is more favorable than a monodispersed system for scattering in a broadband way (Fig. 2E and fig. S8) (*26*). This explains why hierarchical pores render the cooling ceramic not only white in the VIS range, similar to the appearance of *Cyphochilus*, but also white in the UV and NIR ranges.

The cooling ceramic is composed solely of porous α-alumina, which has ideal intrinsic electromagnetic properties for PRC applications. Among common white and transparent inorganic materials, alumina has the top-ranked high bandgap with a relatively high refractive index (fig. S9). The bandgap of α-alumina reaches 7.0 eV, which is well above the upper boundary for photon energy in the solar spectrum (4.13 eV) (*44*, *45*). The high bandgap of α-alumina results in a low extinction coefficient across the solar wavelength range (fig. S10) and a low absorption of solar photons, which makes the high R_{solar} possible. On the other hand, arising from the vibrational modes of the Al–O chemical bonds, alumina attains a high extinction peak within the AW range

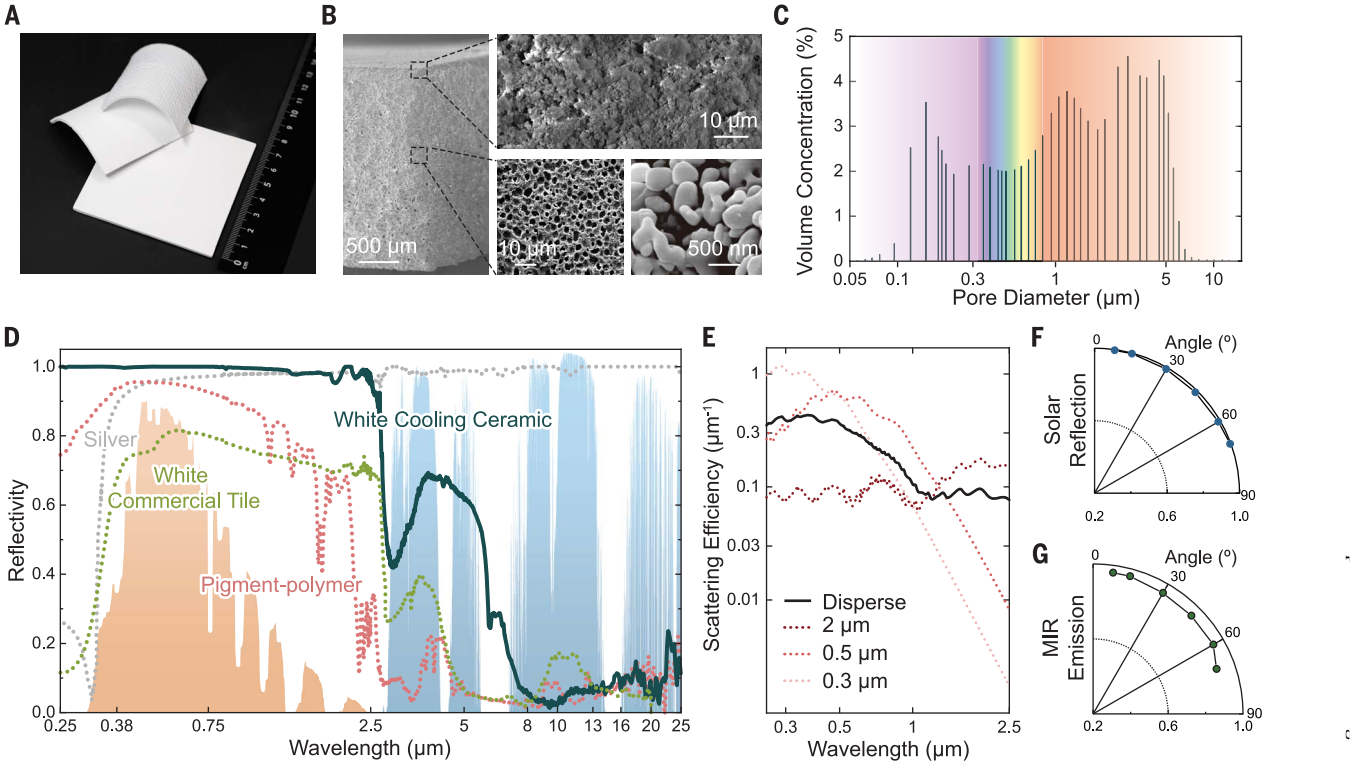


Fig. 2. Engineered hierarchically porous cooling ceramic. (**A**) Photograph of the cooling ceramics with flat and curved shapes. Patterns can also be applied on the cooling ceramic surface. (**B**) SEM images of a fabricated cooling ceramic sample showing the hierarchically porous structure. (**C**) Volume concentration of pores within the cooling ceramic. (**D**) Comparison of the optical properties of the white cooling ceramic with a 600 μm thickness and a 70% porosity to those of white-pigmented polymer (Al_2O_3-doped polydimethylsiloxane), silver, and white commercial tile. (**E**) Simulated scattering efficiency of porous alumina systems, including three monodispersed systems and one multidispersed system (the total pore volume concentrations are the same). The multidispersed case includes 0.3, 0.5, and 2 μm pores with a volume concentration ratio of 1:1:2. (**F**) Angular $R_{solar}(\theta)$ value of the cooling ceramic. (**G**) Angular $\varepsilon_{AW}(\theta)$ value of the white cooling ceramic.

(fig. S10), resulting in a high ε_{AW} of 96.5%. Therefore, the cooling ceramic can efficiently radiate heat through the AW, where peak blackbody emissions from terrestrial surfaces coincide with the high atmospheric transmittance into space. Considering the high R_{solar} and high ε_{AW}, the cooling ceramic can be regarded as superior to other state-of-the-art PRC material counterparts (fig. S11). The whiteness of the cooling ceramic was achieved in a diffused way, with high angular R_{solar} across its full solid angle (Fig. 2F). At the thermal wavelengths, the angular emissivity was also consistent, with only a slight decrease at low zenith angles. We attributed the high ε_{AW} value of the cooling ceramic within the wide angular range (Fig. 2G) to the open, porous surface and the effective medium behavior of the nanopores at high wavelengths (*33*).

The efficient scattering by the hierarchical pore system enables the cooling ceramic to achieve the desired optical performance in a material-saving manner. Cooling ceramic with a thickness of only 150 μm can achieve $R_{solar} > 95\%$, while high-performance roof cooling coatings typically require >1 mm thickness to achieve the same level of optical performance (*46*) (fig. S12). As the thickness of the cooling ceramic increases, its solar transmittance decreases because of increased backscattering, leading to continuous enhancement of R_{solar}. Eventually, when the thickness of the cooling ceramic reaches 600 μm, its R_{solar} reaches saturation at 99.6% (fig. S12). On the other hand, the cooling ceramic exhibits low reflectivity within the AW range at any thickness. Therefore, a thick cooling ceramic can dissipate heat through its own high emissivity, and a thin cooling ceramic does not hinder the radiative dissipation of its substrate, making it favorable for application on substrates with already adequate ε_{AW} values—concrete, for example.

Cooling performance assessment

We demonstrated the cooling performance of our ceramic using a custom-designed thermal setup (*27*) in Hong Kong and compared it to that of commonly used white commercial tiles as a control (R_{solar} = 76.2%, ε_{AW} = 88.5%) (Fig. 3A and table S1). Throughout a continuous 84-hour measurement period, the cooling ceramic consistently maintained a temperature below the ambient air temperature, with an average subambient temperature difference of 3.8°C and a maximum subambient temperature difference of 8.8°C (Fig. 3B and fig. S13). In contrast, the white commercial tiles only provided cooling at night. During the daytime, the cooling ceramic delivered an average subambient temperature reduction of 4.3°C. Given perfect R_{solar}, the cooling ceramic could achieve >4°C subambient cooling even around midday (between 11 a.m. and 2 p.m.), leading to a >8°C lower temperature than that of the white commercial tiles. At the same time, another identical setup coupled with a controllable heating system was installed in parallel to measure the cooling power of the samples. To ensure repeatability, we conducted cooling power measurements over four short periods (Fig. 3C). During the two measurements at night, the cooling ceramic generated average cooling power of 142 and 125 W/m^2, as compared with values of 128 and 117.3 W/m^2 for the white commercial tiles (Fig. 3C). To examine the cooling performance of the cooling ceramic under extreme incoming heat flux conditions, the two daytime measurements were conducted around solar noon, at a solar intensity of up to 800 W/m^2. Despite the challenging conditions, the cooling ceramic demonstrated exceptional cooling, with a cooling power of 134.5 and 133.8 W/m^2 for the two measurements. In contrast, the white commercial tiles, which strongly absorb solar

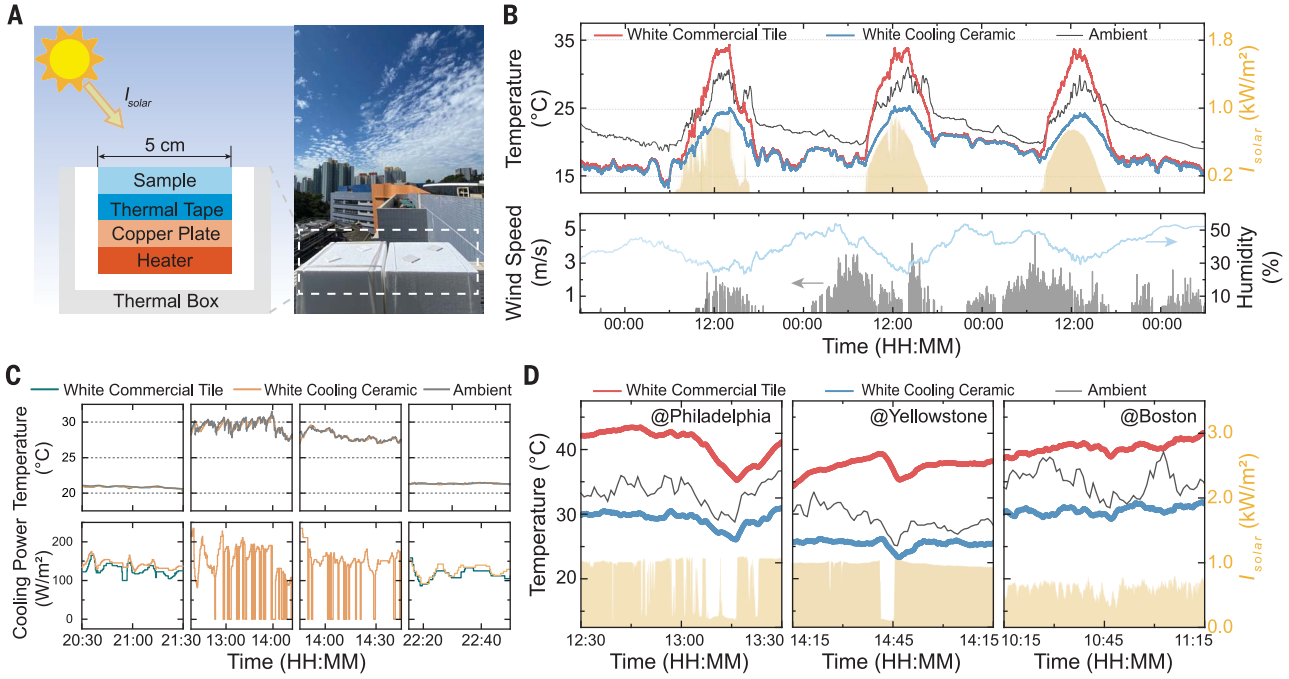


Fig. 3. Cooling performance characterization. (**A**) Experimental setup for characterizing the subambient cooling and cooling power. (**B**) Sample temperatures, ambient air temperature, solar intensity, wind speed, and relative humidity, measured in Hong Kong over 84 continuous hours from 11 to 14 November 2021. (**C**) Cooling power measurements during four periods: 11 November, 20:30–21:30; 12 November, 12:15–14:25; 13 November, 13:45–14:45; and 13 November, 22:15–22:50. (**D**) Temperatures of the samples and ambient air, measured in Philadelphia, Yellowstone National Park, and Boston.

radiation, failed to generate cooling power. For traditional daytime PRC materials, a higher cooling power can normally be obtained at night, while the daytime cooling power is obviously lower than that at night because of solar heat absorption (*22*, *35*). As a result of the ideal R_{solar}, the cooling ceramic exhibited comparable or even superior cooling performance during the daytime as compared with that at night, which is highly desired for addressing the high daytime cooling energy demand. It should be noted that Hong Kong is a coastal city with a humid climate, which is very unfavorable for heat dissipation through thermal radiation. The average humidity during the field test was around 50%. Considering a 3000 atm-cm water column (*47*), the local τ_{AW} during the field test was calculated at <0.6 (*48*), which is not ideal for PRC applications (fig. S1). Although direct comparison to the cooling performance determined in other studies is virtually impossible given the notable influence of geography and meteorological variances on the local field test results, the high cooling power exceeded 100 W/m^2 under these unfavorable test conditions, strongly indicating that the cooling performance of the engineered cooling ceramic is superior to recent results reported in the literature.

Apart from Hong Kong, we also validated the cooling performance of the cooling ceramic in different climates. In Yellowstone National Park (US), the cooling ceramic achieved a cooling effect of subambient temperature reduction of between 1° and 7°C at noon (Fig. 3D). In urban areas, such as Philadelphia (US) and Boston (US), the cooling ceramic also achieved a subambient temperature reduction of about 3°C (Fig. 3D). For the field test conducted in Beijing, the cooling ceramic showed 3.3°C subambient temperature reduction around midday (between 13:30 and 14:30), with an average solar intensity value of 876 W/m^2, whereas the white porous polymer cooler (R_{solar} = 96.2%, ε_{AW} = 96.0%) and white porous alumina-polymer membrane (R_{solar} = 97.6%, ε_{AW} = 97.2%) only showed 2.4° and 2.7°C subambient temperature reduction, respectively; the white commercial tile failed to achieve subambient temperature reduction (fig. S14). These results demonstrate the ability of the cooling ceramic to ensure stable cooling performance in various application environments.

Energy-saving evaluation

External building surfaces are the main recipients of solar radiation and contribute notable heat gains to the indoor environment. The most direct and appealing method to exploit PRC technology would be to use the coolers to cover the building envelope surfaces, where the cooler can be directly exposed to the sky to reduce the building thermal load. To investigate the cooling effect in real applications, two identical model houses were constructed (Fig. 4A). One of them was roofed with white cooling ceramic, while the other was roofed with white commercial tiles purchased from the market as control. The model house roofs were designed with 30° pitch angles to increase the projection areas for direct solar irradiation. First, we conducted continuous thermal measurement for the two model houses over 4 days. The roof temperature showed a large daytime difference between the two model houses (Fig. 4B). Specifically, the roof with white cooling ceramic could be almost 5°C cooler than the commercial tiled roof at noon (Fig. 4B). The lower roof temperature meant reduced heat transfer from the roof to the indoor space, resulting in a lower indoor temperature for the model house fitted with the white cooling ceramic (Fig. 4B). The indoor air temperature difference reached a maximum of 2.5°C. The thermal load of a building is directly related to the cooling demand for regulating the indoor environment. To intuitively quantify the energy-saving potential of the engineered cooling ceramic, we conducted additional assessments by operating an air-conditioning unit within the model houses in the summer (July 2022). We continuously monitored the air-conditioning electricity usage over three periods, with set temperatures of 25°, 23° and 20°C. With less heat load, the cooling ceramic–tiled model house consumed less electricity, with energy savings of 26.8, 22.6, and 19.6%, respectively, during each set temperature period (Fig. 4C). Moreover, we conducted an energy consumption

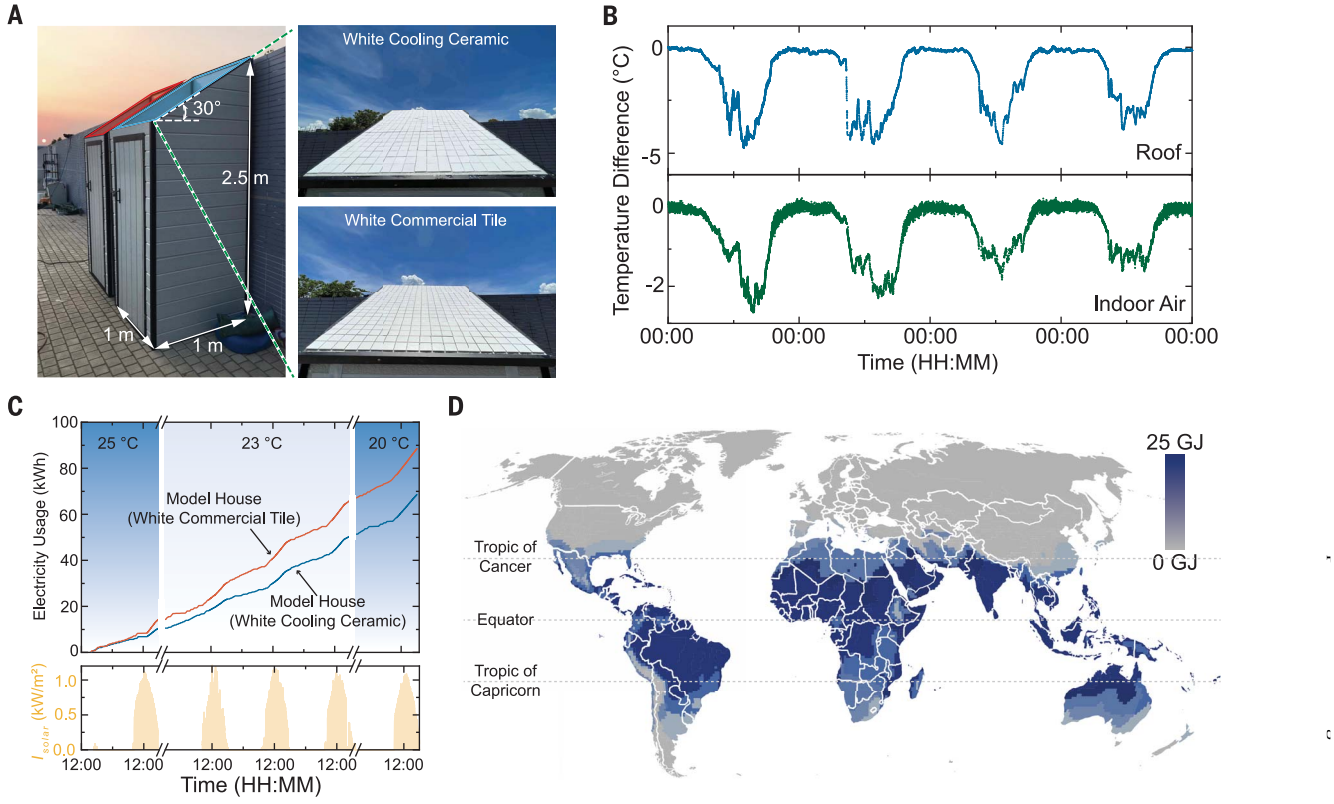


Fig. 4. Application as a building envelope. (**A**) Photograph of the model houses, with the white cooling ceramic and white commercial tile applied on the roof (area: ~1.15 m^2). The model houses were installed 2 m apart during the experiment to eliminate interference. (**B**) Differences in the roof and indoor air temperatures for the two model houses. (**C**) Electricity usage of the two model houses with air-conditioning set points of 25°, 23°, and 20°C. (**D**) Energy-saving performance on a worldwide scale considering the energy consumed by cooling systems, fans, and heating equipment.

simulation of a full-scale building with a model of a typical four-story midrise apartment building (fig. S15 and table S2) to assess the energy-saving performance on a worldwide scale by applying the cooling ceramic as the external envelope material on the walls and roof (Fig. 4D). Considering the energy usage of HVAC (heating, ventilation, and air-conditioning) systems, application of the cooling ceramic could benefit tropical regions the most. In particular, the annual energy savings could reach >10% (25 GJ, or ~7000 kWh) per year for indoor air-conditioning in these extremely hot areas.

Applicability investigation

Beyond cooling, the cooling ceramic could also offer a variety of other functionalities, bringing it closer to practical applications. The interaction between water and surfaces at high temperatures is a critical yet often overlooked phenomenon in the applications involving evaporative cooling. When a building experiences a fire, the heat generated by combustion can cause extensive damage. To extinguish a fire and lower a building's temperature, it is essential to allow direct contact between the envelope surface and water using evaporative cooling. Most commercial tiles prevent water from wetting the overheated surface at temperatures above 280°C (Fig. 5A, fig. S16, and movie S1), resulting in poor evaporative cooling due to the Leidenfrost effect (*49*, *50*). In contrast, our cooling ceramic exhibits superhydrophilicity enabling immediate droplet spreading, while facilitating rapid impregnation of droplets by its interconnected porous structure (fig. S17). Consequently, the cooling ceramic inhibits the Leidenfrost effect at temperatures above 800°C (Fig. 5A and movie S1) during the evaporative cooling process. The effective evaporative cooling performance of the cooling ceramic was demonstrated by the rapidly decreasing surface temperature, whereas only a negligible temperature variation was recorded for the commercial tiles by infrared camera (Fig. 5B and movie S2).

The cooling ceramic can also be converted from superhydrophilic to hydrophobic by impregnation with organosilicon compounds, which is the most commonly used method to protect porous materials against moisture when applications call for water repellence. Upon treatment with fluorosilane, the cooling ceramic featured high superhydrophobicity with a water contact angle of ~150° (fig. S18). The interconnected porous structure of the cooling ceramic enabled the fluorosilane solution to penetrate the material, resulting in water resistance on both the surfaces and interior (fig. S18). Importantly, this treatment does not affect the high-temperature evaporative cooling performance of the cooling ceramic because the fluorosilane chemical groups start to evaporate at a temperature above 250°C (fig. S19). In terms of the optical performance of the fluorosilane-treated cooling ceramic, its R_{solar} dropped slightly to around 99.0% owing to the absorption of bonded fluorosilane chemical groups (fig. S20). A dust resistance test using standard pollutants (GB/T 97780-2013) confirmed good antipollution performance of the fluorosilane-treated cooling ceramic with R_{solar} of >97% maintained after multiple pollutant-washing test cycles (fig. S21).

The cooling ceramic has a dense all-inorganic structure that inherently endows it with excellent resistance to UV radiation. The extremely low UV absorption and high bond strength of alumina makes the cooling ceramic less susceptible than polymer-based coolers to photodegradation. For verification, the cooling ceramic was exposed to a UV lamp with a 5 W/m^2 UV power for 3 months. The optical properties measured before and after UV exposure showed negligible differences, confirming the high resistance of the cooling ceramic to UV degradation (fig. S22). A durability test of the

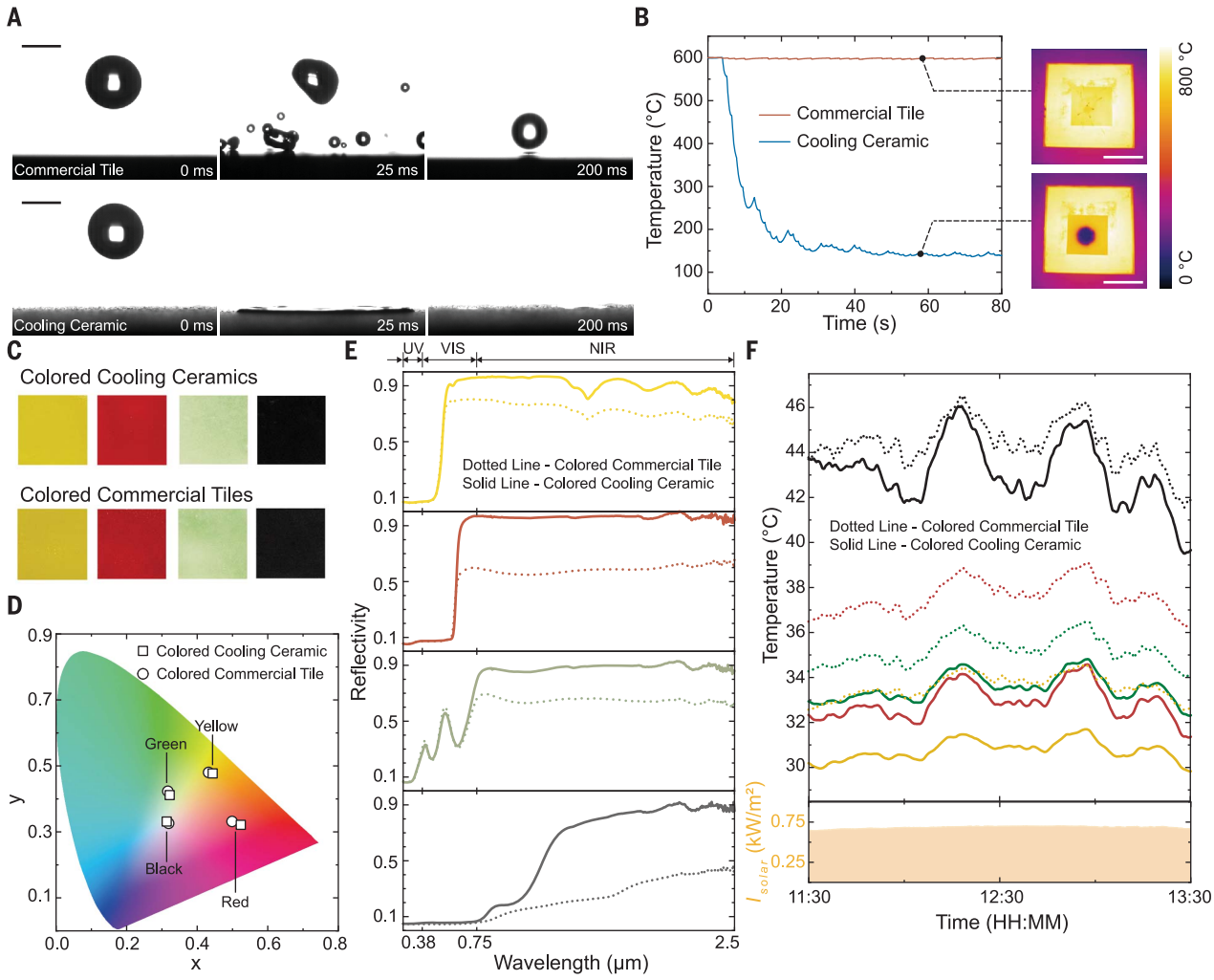


Fig. 5. Applicability of the cooling ceramic. (**A**) Water droplets contacting the commercial tiles and cooling ceramic at a surface temperature of 600°C. With the commercial tiles, water cannot directly contact the surface owing to the Leidenfrost effect. After spreading due to the impact force, the droplets bounce and return to a spherical shape. In contrast, water on the cooling ceramic is constantly pinned and remains flattened upon spreading, leading to rapid evaporation. Scale bars, 2 mm. (**B**) Variation in the surface temperature when the samples are contacted by water droplets at 5-s intervals. The infrared thermal images show the surface temperature of the samples after thermal stabilization. Scale bars, 5 cm. (**C**) Photographs of the fabricated colored cooling ceramics, along with colored commercial tiles showing identical vivid colors. (**D**) Chromaticity of the colored cooling ceramic and colored commercial tiles in the CIE 1931 color space. (**E**) Comparison of the R_{solar} spectra of yellow, red, green, and black cooling ceramics to those of commercialized colored tiles of the same color. (**F**) Outdoor temperature measurements of the colored cooling ceramics around midday, using colored commercial tiles as a control. The line color corresponds to the sample color.

cooling ceramic under real-world conditions was also conducted by exposing it in an outdoor space for 1 year, to solar radiation equaling ~5000 MJ (*51*). The results show that the R_{solar} of the cooling ceramic only dropped by 1.3%, still reaching 98.3% (fig. S23). Because alumina is nonflammable, the cooling ceramic can remain intact and undamaged even under extreme fire exposure (fig. S24). The cooling ceramic also demonstrates high mechanical strength, with a breaking strength of >100 MPa (bending test according to ISO 10545-4), which meets the standards for application as building envelope materials (>35 MPa) (fig. S25). Furthermore, the cooling ceramic can be recycled (fig. S26), which is highly desirable for sustainability. When damaged or polluted, the developed cooling ceramic can be ground into raw materials (alumina particles) and then reused in the fabrication of new cooling ceramic with well-preserved optical properties.

High and broadband R_{solar} is crucial for cooling, yet completely reflecting all visible light would yield a white cooler, which cannot fulfill the aesthetic needs of contemporary urban landscapes. On the basis of the white cooling ceramic, we have developed a colored cooling ceramic that attains equilibrium between chromatic presentation and reduction of thermal load. The colored cooling ceramic was obtained simply by sintering a thin layer of colored glaze on top of the white cooling ceramic. We fabricated four colored cooling ceramics: yellow, red, green, and black. The color of the cooling ceramic was controlled to match that of commercial tile products purchased from the market (Fig. 5, C and D). Both the colored commercial tiles and colored cooling ceramics had high absorption in VIS light for creating vivid color (Fig. 5E and table S3). As the thin color glaze of the colored cooling ceramic enables NIR light to transmit through and be reflected by the underlying highly reflective layer, the colored cooling ceramic attained substantially higher reflectivity in the NIR spectrum (95, 96, 87, and 39% for the yellow, red, green, and black ceramics, respectively) than the colored commercial tiles (76, 57, 66, and 16% for

the yellow, red, green, and black tiles, respectively) (Fig. 5E and table S3). We experimentally verified the thermal performance of the colored ceramics against the colored commercial tiles by exposing the samples to direct sunlight at midday (Fig. 5F). The red cooling ceramic maintained an average temperature 4.7°C lower than that of the commercial red tile, followed by 2.8°, 1.7°, and 1.3°C temperature reductions for the yellow, green, and black cooling ceramics, respectively (Fig. 5F). The temperature reduction obtained for each color corresponded well to the R_{solar} contrast between the cooling ceramic and commercial tiles. The light absorption in the visible range determines the color intensity, which therefore influences the overall R_{solar}. With high NIR reflectivity, the colored cooling ceramic is also able to achieve R_{solar} of >0.9 by tuning the concentration of color glazing (fig. S27). These results confirmed that the colored design based on the white cooling ceramic could provide the desired color along with high NIR R_{solar} and showcased the great potential in reducing the thermal load and even subambient cooling. Finally, the colored cooling ceramics can be fabricated into curved shapes using the same procedure as for the white cooling ceramics. This makes it a versatile material suitable for a wide range of applications, for example, houses with curved roofing tiles (fig. S28). The ability to shape the material into curved forms allows for more creative and distinctive designs, adding aesthetic value to different applications.

Conclusions

In this study, we developed an energy-free and robust daytime PRC material in the form of ceramic for reducing the cooling demand of the indoor environment. The cooler features a biomimetic porous structure, which efficiently scatters solar irradiation to achieve an R_{solar} of 99.6%. In real-world application demonstrations and whole-building energy simulations, our cooling ceramic demonstrated promising energy-saving potential. With high weather resistance, high mechanical strength, favorable recyclability, notable Leidenfrost depression, and optional color features, the cooling ceramic can readily be applied to different scenarios and outdoor infrastructures on a large scale. With its combination of advanced features and demonstrated performance, the technology holds great potential for contributing to the development of more sustainable and energy-efficient building solutions in the future. Although we have not yet explored the cooling power modulation of the cooling ceramic, given its high R_{solar}, the cooling ceramic may integrate with adaptive optical materials to obtain the temperature-response regulation on optical properties and therefore smartly suppress the cooling effect during cold weather (*52*, *53*). On the other hand, to retain a high radiative cooling power with vivid color, we can combine the cooling ceramics with photoluminescent materials (*54*), such as carbon dots, to recover the solar absorption in the visible range.

REFERENCES AND NOTES

1. International Energy Agency (IEA), Space Cooling (IEA Paris, 2022); https://www.iea.org/energy-system/buildings/space-cooling.
2. D. L. Zhao *et al*., *Appl. Phys. Rev.* **6**, 021306 (2019).
3. X. Yin, R. Yang, G. Tan, S. Fan, *Science* **370**, 786–791 (2020).
4. S. H. Fan, W. Li, *Nat. Photonics* **16**, 182–190 (2022).
5. E. Rephaeli, A. Raman, S. Fan, *Nano Lett.* **13**, 1457–1461 (2013).
6. A. P. Raman, M. A. Anoma, L. Zhu, E. Rephaeli, S. Fan, *Nature* **515**, 540–544 (2014).
7. M. M. Hossain, B. H. Jia, M. Gu, *Adv. Opt. Mater.* **3**, 1047–1051 (2015).
8. S. Y. Jeong, C. Y. Tso, Y. M. Wong, C. Y. H. Chao, B. Huang, *Sol. Energy Mater. Sol. Cells* **206**, 110296 (2020).
9. D. G. Baranov *et al*., *Nat. Mater.* **18**, 920–930 (2019).
10. H. Zhang *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 14657–14666 (2020).
11. C. J. Zou *et al*., *Adv. Opt. Mater.* **5**, 1700460 (2017).
12. H. C. Ma *et al*., *Sol. Energy Mater. Sol. Cells* **212**, 110584 (2020).
13. C. Lin *et al*., *Adv. Mater.* **34**, e2109350 (2022).
14. J. L. Kou, Z. Jurado, Z. Chen, S. H. Fan, A. J. Minnich, *ACS Photonics* **4**, 626–630 (2017).
15. S. Y. Jeong *et al*., *Renew. Energy* **146**, 44–55 (2020).
16. D. Lee *et al*., *Nano Energy* **79**, 105426 (2021).
17. Y. Fu, J. Yang, Y. S. Su, W. Du, Y. G. Ma, *Sol. Energy Mater. Sol. Cells* **191**, 50–54 (2019).
18. J. W. Liu *et al*., *Sol. Energy Mater. Sol. Cells* **208**, 110412 (2020).
19. D. Han, B. F. Ng, M. P. Wan, *Sol. Energy Mater. Sol. Cells* **206**, 110270 (2020).
20. C. Y. Tso, K. C. Chan, C. Y. H. Chao, *Renew. Energy* **106**, 52–61 (2017).
21. X. Li, J. Peoples, P. Yao, X. Ruan, *ACS Appl. Mater. Interfaces* **13**, 21733–21739 (2021).
22. X. Y. Li *et al*., *Cell Rep. Phys. Sci.* **1**, 100221 (2020).
23. Z. Tong *et al*., *Mater. Today Phys.* **24**, 100658 (2022).
24. J. Mandal, Y. Yang, N. F. Yu, A. P. Raman, *Joule* **4**, 1350–1356 (2020).
25. X. Xue *et al*., *Adv. Mater.* **32**, e1906751 (2020).
26. J. Peoples *et al*., *Int. J. Heat Mass Transf.* **131**, 487–494 (2019).
27. K. X. Lin *et al*., *Energy Build.* **276**, 112507 (2022).
28. A. Leroy *et al*., *Sci. Adv.* **5**, eaat9480 (2019).
29. H. Zhong *et al*., *ACS Appl. Mater. Interfaces* **12**, 51409–51417 (2020).
30. J. Li *et al*., *Sci. Adv.* **8**, eabj9756 (2022).
31. M. Li *et al*., *Small* **19**, e2301159 (2023).
32. W. L. Huang *et al*., *Adv. Funct. Mater.* **31**, 2010334 (2021).
33. J. Mandal *et al*., *Science* **362**, 315–319 (2018).
34. S. Zeng *et al*., *Science* **373**, 692–696 (2021).
35. T. Li *et al*., *Science* **364**, 760–763 (2019).
36. D. Li *et al*., *Nat. Nanotechnol.* **16**, 153–158 (2021).
37. A. Aili *et al*., *Mater. Today Phys.* **10**, 100127 (2019).
38. J. R. Song *et al*., *Sol. Energy Mater. Sol. Cells* **130**, 42–50 (2014).
39. H. Zweifel, *Stabilization of Polymeric Materials* (Springer Science and Business Media, 2012).
40. S. K. Jeon *et al*., *Adv. Sci. (Weinh.)* **10**, e2302701 (2023).
41. T. Li *et al*., *Chem. Eng. J.* **452**, 139518 (2023).
42. M.-T. Tsai *et al*., *Nano Today* **48**, 101745 (2023).
43. M. Burresi *et al*., *Sci. Rep.* **4**, 6075 (2014).
44. A. Felicelli *et al*., *Cell Rep. Phys. Sci.* **3**, 101058 (2022).
45. E. O. Filatova, A. S. Konashuk, *J. Phys. Chem. C* **119**, 20755–20761 (2015).
46. APOC, Technical and Safety Data Sheets; https://www.apoc.com/pages/technical-data-sheets.
47. Spectral Sciences Inc., MODTRAN (MODerate resolution atmospheric TRANsmission); http://modtran.spectral.com/.
48. C. Ziming *et al*., *AIMS Energy* **9**, 96–116 (2021).
49. D. Quéré, *Annu. Rev. Fluid Mech.* **45**, 197–215 (2013).
50. M. Jiang *et al*., *Nature* **601**, 568–572 (2022).
51. Hong Kong Observatory, Climatological Information Services, Climate of Hong Kong; https://www.hko.gov.hk/en/cis/climahk.htm.
52. S. Wang *et al*., *Science* **374**, 1501–1504 (2021).
53. J. Yang *et al*., *Adv. Mater.* **33**, e2004754 (2021).
54. X. Wang *et al*., *Sci. Bull.* (*Beijing*) **67**, 1874–1881 (2022).

ACKNOWLEDGMENTS

Funding: This work was supported by Hong Kong Research Grant Council via General Research Fund (GRF) accounts 11200022, 11200121, and 11200923 and by the Meituan Green Tech Fund. **Author contributions:** Conceptualization: K.L. and C.Y.T. Methodology: K.L., C.Y.T., X.W., and F.L. Investigation: K.L., S.C., Y.Ze., Y.Zh., and C.Y.-H.C. Visualization: K.L. and C.Y.T. Funding acquisition: C.Y.T. Project administration: C.Y.T. Supervision: C.Y.T. and Z.W. Writing – original draft: K.L. and C.Y.T. Writing – review & editing: C.Y.T., Z.W., B.H., and T.C.H. **Competing interests:** A US nonprovisional patent (application no. 17/89,878) related to this work has been filed. K.L., S.C., Y.Zh., T.C.H., and C.Y.T. are inventors of this filed patent. **Data and materials availability:** All data are available in the main text or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi4725
Materials and Methods
Supplementary Text
Figs. S1 to S28
Tables S1 to S3
References (*55*–*64*)
Movies S1 and S2

Submitted 28 April 2023; accepted 19 September 2023
10.1126/science.adi4725

MAGNETISM

Large effective magnetic fields from chiral phonons in rare-earth halides

Jiaming Luo[1,2], Tong Lin[1], Junjie Zhang[1], Xiaotong Chen[1], Elizabeth R. Blackert[1], Rui Xu[1], Boris I. Yakobson[1], Hanyu Zhu[1]*

Time-reversal symmetry (TRS) is pivotal for materials' optical, magnetic, topological, and transport properties. Chiral phonons, characterized by atoms rotating unidirectionally around their equilibrium positions, generate dynamic lattice structures that break TRS. Here, we report that coherent chiral phonons, driven by circularly polarized terahertz light pulses, polarize the paramagnetic spins in cerium fluoride in a manner similar to that of a quasi-static magnetic field on the order of 1 tesla. Through time-resolved Faraday rotation and Kerr ellipticity, we found that the transient magnetization is only excited by pulses resonant with phonons, proportional to the angular momentum of the phonons, and growing with magnetic susceptibility at cryogenic temperatures. The observation quantitatively agrees with our spin-phonon coupling model and may enable new routes to investigating ultrafast magnetism, energy-efficient spintronics, and nonequilibrium phases of matter with broken TRS.

The term "chiral" most commonly refers to structural chirality without any mirror symmetry in three-dimensional space. Meanwhile, in condensed matter physics, "chiral" sometimes implies a lack of both mirror lines within a two-dimensional plane and time-reversal symmetry (TRS) (*1*–*5*). Chiral wave functions with broken TRS and nonzero angular momenta may have topologically protected properties, such as lossless transport of the chiral edge states in the quantum Hall effect and robust vortices in chiral superconductors. Symmetry breaking is either spontaneous or externally stimulated by magnetic fields, optical excitations, and mechanical motion, which are typically implemented globally on a macroscopic scale (*6*, *7*). The breaking of mechanical TRS may also be implemented at the atomic level and femtosecond time frame, when atoms are displaced away from the equilibrium position inside the lattice and rotate unidirectionally in elliptical trajectories with nonzero angular momenta (*8*–*12*). Such vibrational modes are termed chiral phonons, providing a distinctive approach to controlling the TRS of electronic and magnetic properties. Phononic TRS breaking is a correction to the standard Born-Oppenheimer approximation, which assumes frozen lattices and preserves TRS in solving electronic states. Such correction has been theoretically demonstrated by considering the ionic current (*13*, *14*), electronic Berry phase accumulation (*15*–*18*), and nonadiabatic dynamics (*19*).

Chiral phonons carrying angular momentum have been experimentally validated across multiple material systems and physical processes, including Raman scattering, ultrafast demagnetization, and the thermal Hall effect (*20*–*26*). These phonons have been found to effectively exchange angular momentum with the spin and orbital degrees of freedom (*27*). Occasionally, in materials with strong spin-phonon coupling, the observed phononic effective magnetic moments are orders of magnitude larger than those expected from the ionic loop current (*28*–*30*). Several mechanisms have been proposed to explain such an extraordinary enhancement, including bond-dependent exchange interaction, spin-orbit tilting, hybridization with orbital excitations, and ferroelectric instability (*19*, *31*–*34*). Reciprocal to the field-induced shifts in phonon frequencies, the same mechanism theoretically allows coherent chiral phonons to generate substantial effective magnetic fields inside the materials (*14*, *35*).

However, quantitative studies of phononic magnetism remain elusive because of the challenges in manipulating coherent chiral phonons. Recent advances in nonlinear phononic spectroscopy enabled mode-selective optical excitations of linearly polarized phonons, which have been shown to significantly modulate the structural, electronic, magnetic, and topological properties in many quantum materials (*36*–*40*). In particular, phonon-induced magnetic dynamics have been found by driving one or two linear phonon modes (*40*–*42*). These phonons are not chiral eigenmodes and result in either periodic or impulsive modulation to the interatomic distances, orbital symmetry, and exchange interactions, as opposed to quasi-static breaking of TRS on demand (*43*). Chiral phonons can be excited by strong, narrowband, and circularly polarized (CP) light in the terahertz (THz) frequency range from tabletop nonlinear sou… and free-electron lasers. These excitations … portedly produced ionic Kerr effect and ferromagnetic switching, but the magnetic fields are rather small, estimated to be on the order of millitesla (*44*, *45*).

Chiral phonon–spin coupling in CeF_3

In this study, we quantitatively measured the quasi-static effective magnetic fields that result from chiral phonons in the paramagnetic rare-earth trihalide CeF_3. We found a field strength of >0.9 T from the infrared-active optical phonons centered around 10.5 THz, driven by CP THz pulses with a moderate incident fluence of ~0.4 mJ/cm^2. The effective magnetic field of the phonons polarizes the paramagnetic spin of the Ce^{3+} ions and is quantified by time-resolved Kerr ellipticity. The spin polarization is proportional to the angular momentum of the driving pulses and the phonons, showing the signature of a TRS-breaking dynamic structure. We proved that the sharply increasing magnetization in cryogenic temperatures correlates with the diverging paramagnetic susceptibility, along with the appearance of resonant atomic displacement that is evidenced by time-dependent second-harmonic generation, ruling out the mechanism of a pure transition between the crystal electric field (CEF) levels. From the rate equation of paramagnetic relaxation, we deduced the transient effective magnetic field that quantitatively agrees with the modeled phonon dynamics and the spin-phonon coupling Hamiltonian throughout the temperature range of 10 to 150 K. The strength of the effective magnetic field is proportional to the number of phonons and can potentially reach 50 T under experimentally feasible conditions (*46*). Our method of coherently manipulating magnetic chiral phonons may apply to a broad range of quantum materials to unravel TRS-breaking processes that involve lattice dynamics (*15*–*19*, *31*–*33*).

The magneto-phonon properties of CeF_3 are unusual (fig. S6): The doubly degenerate E_g (391 cm^{-1}) and E_u (~350 cm^{-1}) phonons split in energy when the paramagnetic spins are polarized. At low temperatures, the magnetic susceptibility χ diverges, which enhances the effective Zeeman splitting of these phonons under magnetic fields to >3 cm^{-1}/T, equivalent to a large phononic magnetic moment of >7 μ_B (Bohr magneton) at 1.9 K (*47*–*49*). This moment is about five orders of magnitude larger than what would be expected from ionic current (*13*). It does not saturate at lower temperatures and can exceed the magnetic moments of CEF levels (<2.5 μ_B), meaning that the observed splitting is not simply caused by the hybridization of phonons and CEF levels at different frequencies (*50*–*52*). Rather, a more plausible mechanism might be the spin-dependent

[1]Department of Materials Science and Nano Engineering, Rice University, Houston, TX 77005, USA. [2]Applied Physics Graduate Program, Rice University, Houston, Texas 77005, USA.
*Corresponding author. Email: hanyu.zhu@rice.edu

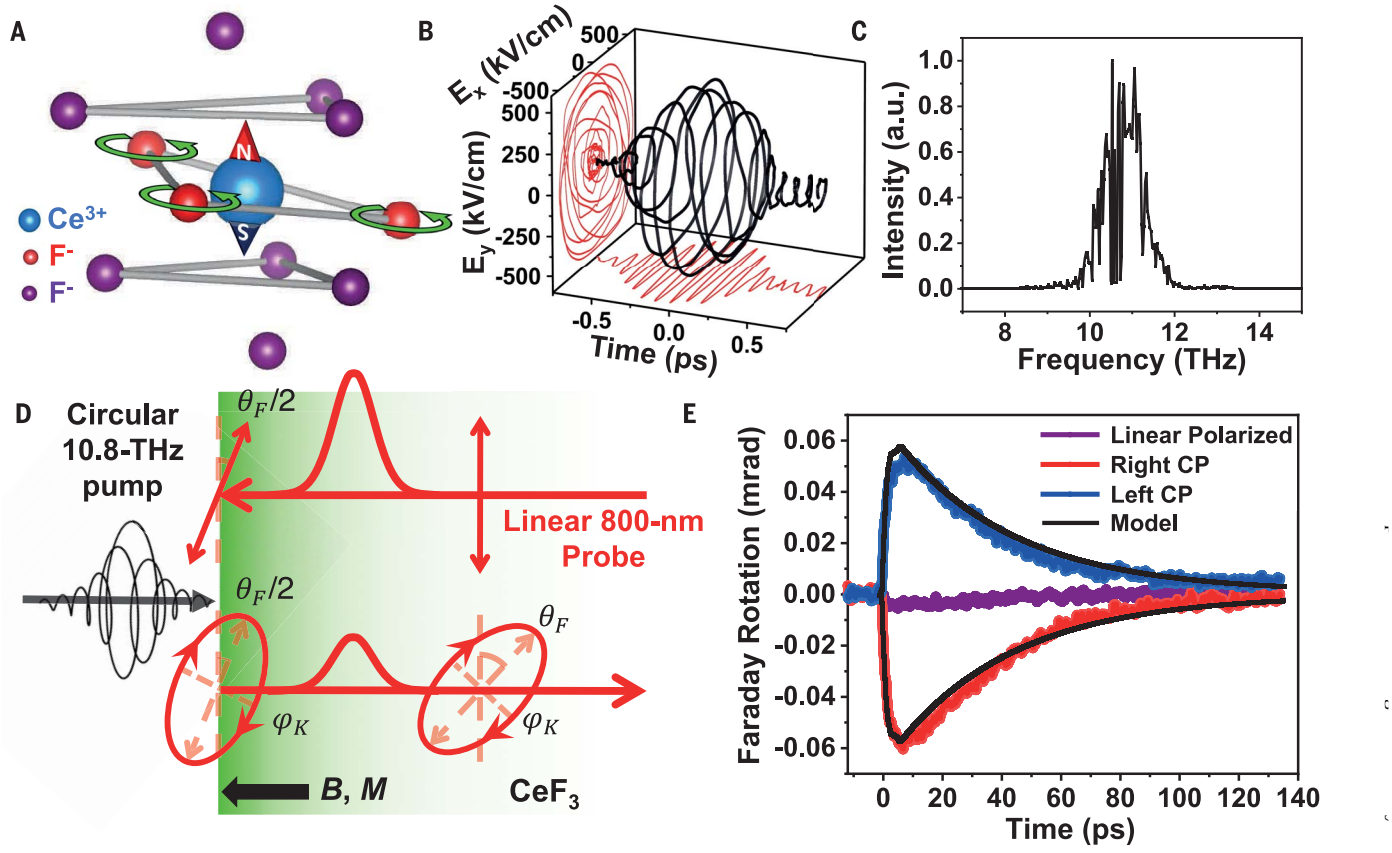


Fig. 1. Ultrafast magnetization induced by chiral phonons in CeF_3. (**A**) Illustration of part of the unit cell centered around a Ce^{3+} ion and the atomic displacement of E_u chiral phonon mode centered around 10.5 THz according to DFT calculations. The F^- ions are denoted in red and purple according to their difference in effective charge tensor. (**B**) Electric field of the CP THz pump pulse (with a helicity of 88%) as a function of time measured by electro-optic sampling. The peak field incident on the material is ~5.6 × 10^7 V/m, and the pulse duration is ~0.45 ps. (**C**) In the frequency domain, the pulse is centered at 10.8 THz with 1-THz bandwidth. (**D**) Schematic of the time-resolved Kerr ellipticity and Faraday rotation process. The CP THz pump pulses are incident from the left and exponentially decay inside the crystal, which generate an effective magnetic field perpendicular to the surface and a layer of transient magnetization (green shade) on the order of 1 μm in thickness. The linearly polarized probe with a wavelength centered at 800 nm and a duration of 0.5 ps comes from the back of the sample, experiencing twice the Faraday rotation from the magnetization layer, as well as the Kerr ellipticity from reflectance at the interface. (**E**) Faraday rotation as a function of delay time is opposite under left CP and right CP THz excitation, evidencing TRS breaking.

renormalization of phonon frequencies by nonresonant CEF (*53*), which has reproduced the order of magnitude of spin-phonon coupling in $CeCl_3$ using a simple point-charge model (*34*). Without knowing the microscopic theories, a simple phenomenological Hamiltonian under the symmetry constraint is sufficient to connect the observed phonon magnetic moment with the phonon-induced effective magnetic field: $H = K\boldsymbol{\mu} \cdot \boldsymbol{L}$, where $\boldsymbol{\mu}$ and $\boldsymbol{L}$ are the spin magnetic moment and the net angular momentum of phonons, respectively, and K is proportional to the phononic magnetic moment (*35*). When the spins are fully polarized to the saturation magnetic moment $\boldsymbol{\mu}_S$ by external fields, the frequency of a chiral phonon mode shifts by a saturation value $\Delta\Omega = 0.3 \pm 0.1$ THz. Conversely, when chiral phonons are present, the energy of the spins shifts as if influenced by an effective magnetic field [i.e., a phononic inverse Faraday effect (*54*)]:

$$B_{\text{eff}} = \frac{\hbar\Delta\Omega}{\mu_S} n \quad (1)$$

THz phonon-induced magnetization

In our experiment, we excite chiral phonons in c-cut CeF_3 by using normal-incident, resonant, and CP THz pump pulses. The doubly degenerate Raman- and infrared-active modes, which are enabled by the trigonal $\bar{P}3C1$ space group of the lattice, can both be expressed in CP basis along the c axis with a pseudo-angular momentum quantum number ±1 (table S2). Among the 12 possible pairs of chiral phonons, we chose to measure the effective magnetic field of the infrared-active, doubly degenerate phonon pairs $E_u^{\pm}$ centered around 10.5 THz, because of their large angular momentum according to density functional theory (DFT) calculations and the most prominent magnetic field–dependent infrared activity. The displacements of all six Ce^{3+} ions in the unit cell are in phase for these modes, so for clarity, we only plot a part of the unit cell in Fig. 1A. The F^- ions in the Ce^{3+} plane exhibit the largest displacement, modifying the local crystal field and mixing the CEF levels under the constraint of pseudo-angular momentum conservation. The other F^- ions are in equivalent crystallographic positions and have small displacement. The $E_u^{\pm}$ phonons are selectively and strongly coupled with CP THz photons, which are created by chirped-pulse difference frequency generation in nonlinear organic crystals and a variable phase delay between two cross-linearly polarized beams (fig. S1) (*54*). The maximum helicity of the combined THz pulses reaches ~88% (Fig. 1B and fig. S10).

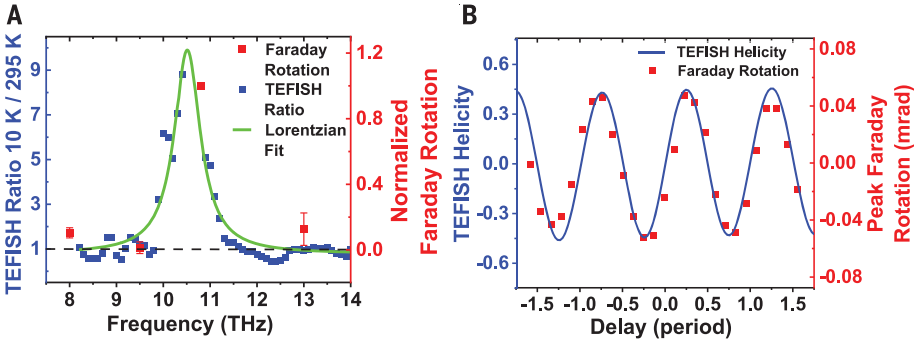


Fig. 2. Correlation between the coherent chiral phonons and the magnetization. (**A**) Ratio between the TEFISH spectra measured at 10 and 295 K. A Lorentzian fit reveals a resonant phonon peak centered around 10.5 THz. The same resonance is observed for the Faraday rotation induced by THz excitation at different frequencies, normalized by the pulse energy, confirming that the observed magnetization has a phononic origin. (**B**) Calculated helicity of the TEFISH signal and the measured peak Faraday rotation at 10 K as a function of retardation (in the unit of phononic period) between the horizontally and vertically polarized THz pump pulses, demonstrating that the magnetization arises from the angular momentum of phonons.

We then measure the phonon-induced magnetization by time-resolved magneto-optic Kerr ellipticity and Faraday rotation (Fig. 1C) (*54*). The linearly polarized near-infrared probe pulses are tightly focused onto the center of the THz pump pulses by an objective lens from the back of the sample. The pulse duration is ~0.5 ps because of the dispersion of the optics, which limits the temporal resolution, but at the same time filters out irrelevant, fast-oscillating field-induced polarization effects. The penetration depth of the THz field and the layer of magnetization is <1 μm thick, so group velocity mismatch between the pump and probe fields is negligible. Because CeF_3 is nearly lossless at the probe wavelength, the Kerr ellipticity mainly comes from the reflectance at the vacuum-sample interface, and the Faraday rotation mainly comes from transmitting through the entire layer of magnetization. The two effects happen to have comparable magnitude in our experiment and correspond to the same magnetization dynamics (figs. S4 and S8). Figure 1E shows the Faraday rotation of CeF_3 at 10 K excited by left CP, right CP, and linearly polarized THz pulses near the phonon resonance. The spin accumulation stage lasts a few picoseconds, longer than the pulse duration, and cannot be explained by the optical inverse Faraday effect or the direct optical excitation of the CEF levels. Instead, additional degrees of freedom must be present, such as phonons, magnons, or metastable states, to mediate the light-spin interaction (*39*). The magnetization switches sign for pulses with opposite helicity and approaches zero when TRS is preserved under linearly polarized excitation. The remnant signal comes from a small THz ellipticity caused by the birefringence of the diamond window. Such TRS dependence means that the magnetization dynamics is not from thermal or doping effects.

Next, we prove that coherent chiral phonons are the source of the observed magnetization by heterodyne THz electric field–induced second-harmonic generation (TEFISH) (*54*, *55*). Normally, TEFISH in thin films is dominated by atomic displacement when the THz field is resonant with the phonons (*55*). Yet in bulk crystals, the contribution from pure electric field effect is magnified by phase matching and difficult to separate from the atomic contribution at room temperature (fig. S9). Fortunately, the atomic contribution grows with the phononic coherence lifetime, which increases at cryogenic temperatures, and thus can be extracted from the temperature dependence of TEFISH in bulk CeF_3. Figure 2A shows the ratio of frequency-domain TEFISH signals measured at 10 K and room temperature. We found that the ratio is close to 1, suggesting a temperature-independent field effect, in a broad range of 8 to 14 THz except for a sharp resonance centered around 10.5 THz, which agrees with the E_u phonons measured by ellipsometry (fig. S7). Therefore, we attribute this resonant component to the atomic displacement from coherent phonons. By varying the center frequency of the THz excitation, we found that the magnetization only occurs near the phonon resonance and thus cannot be attributed to the inverse Faraday effect from the THz electric field. Furthermore, we verified that the maximum magnetization occurs at atomic displacements with almost equal amplitude in the horizontal and vertical directions but a $T/4$ shift in time (i.e., when the atoms are rotating), where T is the period of the phonon mode (fig. S11). We continuously tuned this shift and measured the peak magnetization in comparison with the helicity of the TEFISH signal (Fig. 2B). The magnetization and helicity oscillate together, which confirms that the observed magnetization is proportional to the phonon angular momentum. The loss of helicity from the free-space THz pulses to the TEFISH signal is largely caused by the phase difference between the atomic and field contributions (*54*).

Modeling the spin and phonon dynamics

In addition, a detailed analysis of the temperature-dependent spin dynamics also validates the mechanism of phonon-induced paramagnetic relaxation. Direct pulsed THz excitation of spin-orbit transition between CEF levels should initiate constant spin polarization, but we found that the magnetization is very weak at temperatures >150 K compared with that measured at 10 K (Fig. 3A). The magnetization M is converted from the observed Kerr ellipticity with temperature- and wavelength-dependent Verdet constant (eq. S20) from literature (*56*–*58*). The greatly enhanced magnetization at cryogenic temperatures can be attributed to two factors. First, a sharper phonon resonance and longer phonon lifetime at lower temperatures lead to a larger phonon population and a stronger, longer-lasting effective magnetic field. Second, the diverging magnetic susceptibility results in more spin polarization, despite slower spin response to the field. To quantify this process, we derived the time-dependent effective magnetic field from the measured paramagnetic relaxation (*54*):

$$\frac{dM}{dt} = \frac{\chi B_{\text{eff}} - M}{\tau_{\text{spin}}} \tag{2}$$

Considering the effective magnetic field B_{eff} is short-lived, we fit τ_{spin} by the relaxation of magnetization and then calculate B_{eff} from $M(t)$ and τ_{spin}. In parallel, we solved the coherent phonon field $|Q\rangle$ as a driven mechanical oscillator by the THz field $\vec{\boldsymbol{E}}$ inside the material experienced by the phonons:

$$\frac{d^2Q}{dt^2} + \frac{1}{\tau_{\text{phonon}}}\frac{dQ}{dt} + \Omega^2 Q = \sum_i \frac{\Omega}{2h}\vec{\boldsymbol{E}} \cdot \boldsymbol{e}_i^* \cdot \vec{\boldsymbol{A}}_i \tag{3}$$

where $\boldsymbol{e}_i^*$ is the Born effective charge tensor of the ith atom in the unit cell, and $\vec{\boldsymbol{A}}_i$ is the displacement in a single chiral phonon, both calculated from DFT. The effective magnetic field is then calculated from Eq. 1 by using $n = |Q|^2$ and considering the imperfect helicity of the phonons (*54*). Comparing the measured B_{eff} with that from the phonon model (Fig. 3B), we can fit the phonon lifetime. Figure 3C shows the summary of the spin and phonon lifetimes from 10 to 150 K. The spin lifetime drops from 39 ps at 10 K to 2 ps at 150 K, and the phonon lifetime decreases from 0.6 ps at 10 K to ~0.1 ps

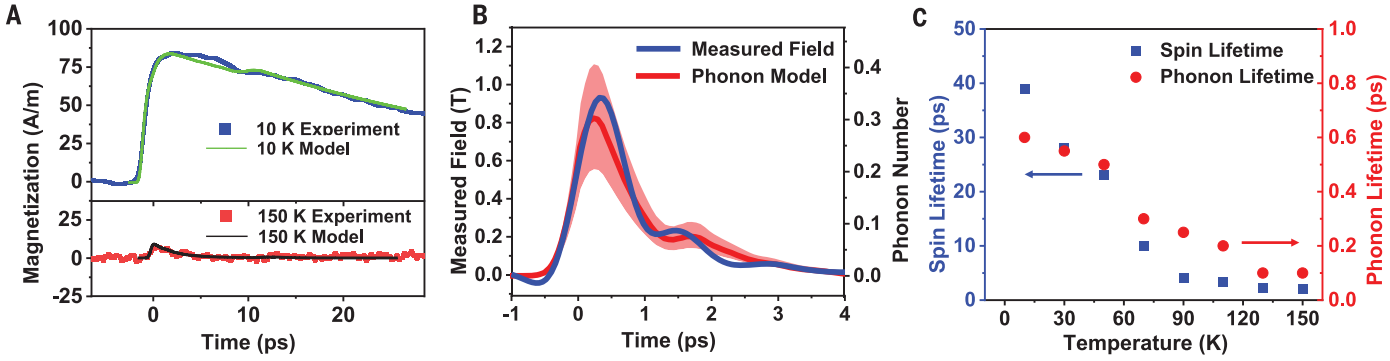
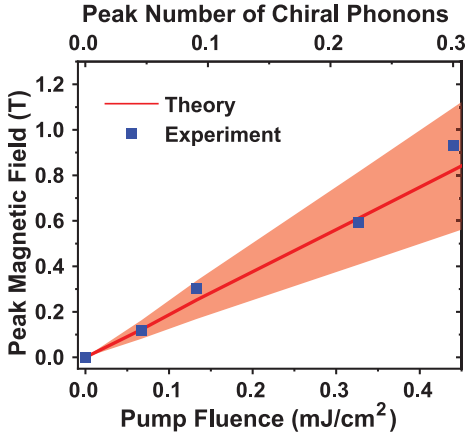


Fig. 3. The matching dynamics of magnetization and chiral phonons. (**A**) Comparing the time-resolved magnetization at 150 K, the much larger magnetization at 10 K agrees with paramagnetic relaxation in a phononic effective magnetic field and rules out direct spin excitation. (**B**) Time-dependent effective magnetic field derived from measured magnetization dynamics at 10 K aligns with that derived from phonon dynamics, using spin and phonon lifetimes as the only fitting parameters. The shadowed region indicates uncertainty from the measurement of $\Delta\Omega$. (**C**) Spin and phonon lifetimes obtained at 10 to 150 K, which reproduce the magnetization dynamics in (A) with the models of Eqs. 1 and 3.

Fig. 4. Scaling of chiral phonon–induced effective magnetic field. The experimental peak field is proportional to the incident fluence of THz excitation, as well as the peak number of chiral phonons per unit cell, as expected from theoretical Eq. 1. The shadowed region indicates uncertainty from the measurement of $\Delta\Omega$.



at 150 K, consistent with values obtained in similar materials (*59*). With only these two lifetimes and no free parameter to adjust the magnitude, we reproduced the magnetization throughout the temperature range (Fig. 3A and figs. S13 and S14), which validates our model of spin-phonon coupling (Eq. 1).

Finally, we verify that the effective magnetic field scales linearly with the chiral phonon population, in accordance with the symmetry requirement of spin-phonon coupling (Fig. 4). Under varying pump fluence, we measured the corresponding peak magnetization and calculated the chiral phonon population, and the slope agrees with the theoretical expectation. The maximum field strength achieved in our experiment is 0.93 T under a moderate fluence of 0.44 mJ/cm^2, corresponding to a net absorbed fluence of <0.2 mJ/cm^2 and a bond deformation of 0.7% (fig. S12). The relatively low fluence used in our experiment ensures that the magnetization is unlikely to be caused by nonlinear effects such as phonon anharmonicity or lattice-induced phase transition (*36*, *60*). The linear trend suggests that the transient effective magnetic field could approach 50 T when the absorbed fluence is >10 mJ/cm^2 (*39*, *46*), which stays below the Lindemann melting criterion and avoids significant anharmonicity (*61*).

Conclusions

We have experimentally observed the chiral phonon–induced magnetization that corresponds to effective magnetic fields on the order of 1 T in CeF_3. The magnetization of the material is controlled by the helicity of incident THz excitation and the phonons and is quantified by time-resolved magneto-optic spectroscopy. We firmly established the phononic origin of the transient magnetization by frequency- and temperature-dependent measurements, which correlate with the second-harmonic generation from the phononic structural symmetry breaking. We elucidated the pathway of angular momentum transfer from CP THz light to spins through chiral phonons and ruled out other alternative mechanisms, including the photonic inverse Faraday effects and the excitation of CEF levels. These time-resolved spin dynamics are quantitatively explained by a phenomenological spin-phonon coupling Hamiltonian and the rate equations of phonons and spins, with only two free parameters of phonon and spin lifetimes. The magnetic chiral phonons offer a new route to coherent engineering of quantum materials and THz spintronic devices.

REFERENCES AND NOTES

1. X. G. Wen, F. Wilczek, A. Zee, *Phys. Rev. B Condens. Matter* **39**, 11413–11423 (1989).
2. S. Raghu, F. D. M. Haldane, *Phys. Rev. A* **78**, 033834 (2008).
3. X.-L. Qi, S.-C. Zhang, *Rev. Mod. Phys.* **83**, 1057–1110 (2011).
4. L. Zhang, Q. Niu, *Phys. Rev. Lett.* **115**, 115502 (2015).
5. S.-W. Cheong, *Npj Quantum Mater.* **4**, 1–9 (2019).
6. R. Fleury, D. L. Sounas, C. F. Sieck, M. R. Haberman, A. Alù, *Science* **343**, 516–519 (2014).
7. J. Del Pino, J. J. Slim, E. Verhagen, *Nature* **606**, 82–87 (2022).
8. H. Zhu *et al.*, *Science* **359**, 579–582 (2018).
9. H. Chen, W. Zhang, Q. Niu, L. Zhang, *2D Mater.* **6**, 012002 (2018).
10. Z. Li *et al.*, *ACS Nano* **13**, 14107–14113 (2019).
11. E. Liu *et al.*, *Phys. Rev. Res.* **1**, 032007 (2019).
12. M. He *et al.*, *Nat. Commun.* **11**, 618 (2020).
13. D. M. Juraschek, N. A. Spaldin, *Phys. Rev. Mater.* **3**, 064405 (2019).
14. R. M. Geilhufe, V. Juričić, S. Bonetti, J.-X. Zhu, A. V. Balatsky, *Phys. Rev. Res.* **3**, L022011 (2021).
15. C. Xiao, Y. Ren, B. Xiong, *Phys. Rev. B* **103**, 115432 (2021).
16. Y. Ren, C. Xiao, D. Saparov, Q. Niu, *Phys. Rev. Lett.* **127**, 186403 (2021).
17. L.-H. Hu, J. Yu, I. Garate, C.-X. Liu, *Phys. Rev. Lett.* **127**, 125901 (2021).
18. G. Xiong, H. Chen, D. Ma, L. Zhang, *Phys. Rev. B* **106**, 144302 (2022).
19. D. Shin *et al.*, *Nat. Commun.* **9**, 638 (2018).
20. A. S. Pine, G. Dresselhaus, *Phys. Rev.* **188**, 1489–1496 (1969).
21. G. Grissonnanche *et al.*, *Nat. Phys.* **16**, 1108–1111 (2020).
22. S. G. Jeong *et al.*, *Sci. Adv.* **8**, eabm4005 (2022).
23. S. R. Tauchert *et al.*, *Nature* **602**, 73–77 (2022).
24. K. Ishito *et al.*, *Nat. Phys.* **19**, 35–39 (2022).
25. W. J. Choi *et al.*, *Nat. Photonics* **16**, 366–373 (2022).
26. K. Kim *et al.*, *Nat. Mater.* **22**, 322–328 (2023).
27. B. Koopmans *et al.*, *Nat. Mater.* **9**, 259–265 (2010).
28. G. Schaack, *Solid State Commun.* **17**, 505–509 (1975).
29. B. Cheng *et al.*, *Nano Lett.* **20**, 5991–5996 (2020).
30. A. Baydin *et al.*, *Phys. Rev. Lett.* **128**, 075901 (2022).
31. B. Flebus, A. H. MacDonald, *Phys. Rev. B* **105**, L220301 (2022).
32. X.-Q. Sun, J.-Y. Chen, S. A. Kivelson, *Phys. Rev. B* **106**, 144111 (2022).
33. R. M. Geilhufe, *Phys. Rev. Res.* **4**, L012004 (2022).
34. S. Chaudhary, D. M. Juraschek, M. Rodriguez-Vega, G. A. Fiete, arXiv.2306.11630 (2023).

35. D. M. Juraschek, T. Neuman, P. Narang, *Phys. Rev. Res.* **4**, 013129 (2022).
36. M. Först *et al*., *Nat. Phys.* **7**, 854–856 (2011).
37. B. Liu *et al*., *Phys. Rev. X* **10**, 011053 (2020).
38. A. S. Disa, T. F. Nova, A. Cavalleri, *Nat. Phys.* **17**, 1087–1092 (2021).
39. A. S. Disa *et al*., *Nat. Phys.* **16**, 937–941 (2020).
40. A. Stupakiewicz *et al*., *Nat. Phys.* **17**, 489–492 (2021).
41. T. F. Nova *et al*., *Nat. Phys.* **13**, 132–136 (2017).
42. D. Afanasiev *et al*., *Nat. Mater.* **20**, 607–611 (2021).
43. D. N. Basov, R. D. Averitt, D. Hsieh, *Nat. Mater.* **16**, 1077–1088 (2017).
44. M. Basini *et al*. arXiv:2210.01690 (4 October 2022).
45. C. S. Davies, N. Fennema, A. Tsukamoto, A. Kirilyuk, "Ultrafast Helicity-Dependent Magnetic Switching by Optical Phonons Driven at Resonance" in *2022 47th International Conference on Infrared, Millimeter and Terahertz Waves (IRMMW-THz)* (2022), p. 1.
46. M. Seo, J.-H. Mun, J. Heo, D. E. Kim, *Sci. Rep.* **12**, 16273 (2022).
47. G. Schaack, *J. Phys. C Solid State Phys.* **9**, L297–L301 (1976).
48. G. Schaack, *Physica B+C* **89**, 195–200 (1977).
49. H. Gerlinger, G. Schaack, *J. Phys. Colloq.* **42**, C6-499, C6–C501 (1981).
50. J. M. Baker, R. S. Rubins, Electron Spin Resonance in Two Groups of Lanthanon Salts, *Proc. Phys. Soc.* **78**, 1353–1360 (1961).
51. R. P. Leavitt, C. A. Morrison, *J. Chem. Phys.* **73**, 749–757 (1980).
52. H. Gerlinger, G. Schaack, *Phys. Rev. B Condens. Matter* **33**, 7438–7450 (1986).
53. P. Thalmeier, P. Fulde, *Z. Phys. B Condens. Matter Quanta* **26**, 323–328 (1977).
54. Materials, methods, and additional information are available in the supplementary materials.
55. T. Lin *et al*., Optica Open [Preprint] (2023); https://doi.org/10.1364/opticaopen.22066295.v2.
56. C. Leycuras, H. Le Gall, M. Guillot, A. Marchand, *J. Appl. Phys.* **55**, 2161–2163 (1984).
57. L. Gong-qiang, Z. Wen-kang, Z. Xing, *Phys. Rev. B Condens. Matter* **48**, 16091–16094 (1993).
58. D. Vojna *et al*., *Opt. Eng.* **56**, 067105 (2017).
59. R. Kolesov *et al*., *Phys. Rev. Lett.* **111**, 120502 (2013).
60. A. Subedi, A. Cavalleri, A. Georges, *Phys. Rev. B Condens. Matter Mater. Phys.* **89**, 220301 (2014).
61. A. von Hoegen, R. Mankowsky, M. Fechner, M. Först, A. Cavalleri, *Nature* **555**, 79–82 (2018).

ACKNOWLEDGMENTS

We thank A. Donohue from J.A. Woollam Co., Inc., for assistance in measuring the dielectric function of the CeF_3 crystal. We also thank D. Juraschek (Tel Aviv University) and S. Zollner (New Mexico State University) for helpful discussions. **Funding:** National Science Foundation grants DMR-2005096 (J.L., X.C., and H.Z.), 1842494 (E.R.B.), and DMR-2240106 (H.Z.); Welch Foundation grant C-2128 (R.X., T.L., and H.Z.); Army Research Office grant W911NF-16-1-0255 (J.Z. and B.I.Y.); and Office of Naval Research grant N00014-22-1-2753 (J.Z. and B.I.Y.). **Author contributions:** Conceptualization: H.Z.; Methodology: J.L., X.C., and H.Z.; Investigation: J.L., T.L., J.Z., and H.Z.; Validation: J.L. and H.Z.; Visualization: J.L., E.R.B., R.X., and H.Z.; Funding acquisition: H.Z. and B.I.Y.; Writing: J.L., T.L., J.Z., X.C., E.R.B., R.X., B.I.Y., and H.Z. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions are available in the main text or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi9601
Materials and Methods
Supplementary Text
Figs. S1 to S14
Tables S1 and S2
References (*62*–*76*)

Submitted 1 June 2023; accepted 7 September 2023
10.1126/science.adi9601

WILDFIRES

Rising wildfire risk to houses in the United States, especially in grasslands and shrublands

Volker C. Radeloff[1]*, Miranda H. Mockrin[2], David Helmers[1], Amanda Carlson[1]†, Todd J. Hawbaker[3], Sebastian Martinuzzi[1], Franz Schug[1], Patricia M. Alexandre[1]‡, H. Anu Kramer[1], Anna M. Pidgeon[1]

Wildfire risks to homes are increasing, especially in the wildland-urban interface (WUI), where wildland vegetation and houses are in close proximity. Notably, we found that more houses are exposed to and destroyed by grassland and shrubland fires than by forest fires in the United States. Destruction was more likely in forest fires, but they burned less WUI. The number of houses within wildfire perimeters has doubled since the 1990s because of both housing growth (47% of additionally exposed houses) and more burned area (53%). Most exposed houses were in the WUI, which grew substantially during the 2010s (2.6 million new WUI houses), albeit not as rapidly as before. Any WUI growth increases wildfire risk to houses though, and more fires increase the risk to existing WUI houses.

Wildfires pose substantial risks to people and their homes (*1*), and that risk is concentrated in the wildland-urban interface (WUI), where wildland vegetation and houses are in close proximity (*2*, *3*). Concerns about wildfire risk have grown rapidly in the US over the past several decades, with more than 55,000 houses burned from 2010 to 2022 (*4*) as a result of rapid increases in both the area that burns annually (*5*, *6*) and the number of houses in the WUI (*7*). However, despite widespread concerns about wildfire risk, patterns of wildfire extent and building exposure in different types of vegetation and WUI are unknown. In this work, we drew upon three decades of data on wildfire occurrence and WUI growth in the US, from 1990 to 2020, to examine these trends.

Wildland vegetation type greatly affects fire behavior (*8*). Crown fires in forests have the highest fire intensity and can produce many embers (*9*) that can ignite houses that are far from a fire front. However, grassland and shrubland fires can spread rapidly when wind is strong, such as in the 2021 Marshall Fire near Boulder, Colorado, which destroyed more than a thousand houses. Furthermore, fuels recover quickly in grasslands and shrublands, such that areas can reburn within a few years, and these areas require different risk management strategies than forests (*10*). Our first goal was to assess which vegetation types dominated (i) fires, (ii) the vicinity of buildings destroyed by fires, and (iii) the WUI at-large.

[1]SILVIS Lab, Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, Madison, WI 53706, USA. [2]Northern Research Station, US Department of Agriculture Forest Service, Catonsville, MD 21228, USA. [3]US Geological Survey, Geosciences and Environmental Change Science Center, Lakewood, CO 80225, USA.
*Corresponding author. Email: radeloff@wisc.edu
†Present address: US Geological Survey, Geosciences and Environmental Change Science Center, Lakewood, CO 80225, USA.
‡Present address: Land and Water Resources Department, County of Dane, Madison, WI 53703, USA.

The exposure of houses to fires represents realized risk and can rise either as a result of housing growth or increases in burned area. Over the past 50 years, housing growth has far outpaced population growth in the US, owing to smaller household sizes and more second homes (*11*). Concomitantly, burned area has increased substantially (*5*, *6*) as a result of fuel accumulation (*12*, *13*), human ignitions (*14*, *15*), and, in the West, earlier spring snowmelt and extended droughts (*16*, *17*). Our second goal was thus to determine (i) how much the exposure of houses to fires has grown, (ii) whether that increase was due mainly to housing growth or an increase in burned area, and (iii) how many houses were built within fire perimeters after the fires occurred.

Although many houses have already been exposed to fires, many more are at risk, which is why WUI growth at-large is a concern. Indeed, during the 1990s and 2000s, the WUI was the fastest-growing land cover type in the US (*7*). However, housing construction plummeted after the 2008 economic recession, and an increase in fires may have slowed WUI growth if homeowners decided to build in non-WUI areas in response or local planners limited development in the WUI. Our third goal was to (i) quantify overall growth in WUI areas and houses, (ii) compare regional patterns, and (iii) assess the effects of the 2008 housing market downturn.

Vegetation type within wildfire perimeters and in the WUI

News coverage of wildfires and discussions about fire-management efforts, such as fuel treatments, tend to focus on forest fires (*18*, *19*). However, we found that within the perimeters of the wildfires that burned from 1990 to 2020 in the conterminous US, grassland and shrubland covered 64.0% of the area burned (33.7 million ha), whereas only 27.3% was forest (14.4 million ha) and 8.7% was other land cover (4.6 million ha) (Fig. 1A) (*20*).

Wildfires are most destructive in the W[illegible] of which there are two types: intermix [illegible] interface. Intermix WUI is the type in which houses directly intermingle with wildland vegetation, and so vegetation has a more direct effect on wildfire risk to houses because many are ignited by nearby burning vegetation (*21*). In the intermix WUI areas that burned, the majority burned as a result of grassland and shrubland fires (44.5 versus 41.2% as a result of forest fires; Fig. 1B) but not by as wide a margin as for the entire wildfire area (Fig. 1A). Interface WUI contains less wildland vegetation than intermix WUI but is in close proximity to a large wildland vegetation area. Although wildfires do destroy many houses in interface WUI (*8*), they are typically ignited by fire brands originating from afar or when nearby houses burn, and so we present vegetation results for interface WUI in the supplementary materials (fig. S6).

In the entire intermix WUI, not just in the intermix WUI that burned, grasslands and shrublands were also widespread (Fig. 1C). Especially in those western states where wildfires are most prevalent, grasslands and shrublands dominate the intermix WUI (Fig. 1G). For example, in California, 52.3% of the intermix WUI is grassland and shrubland and only 30.1% is forest. In Colorado, grasslands and shrublands also dominate (45.7 versus 35.2% forest), and in Arizona, shrublands alone cover 74.1% of the intermix WUI. However, across the US, forest is the most common wildland vegetation type in intermix WUI, covering a little more than half of the area (53.8%; Fig. 1C), because grasslands and shrublands are uncommon in the large intermix WUI areas of the eastern US. Over time, grasslands and shrublands dominated burned areas in most years from 1990 to 2020 (Fig. 1D), and also the intermix WUI that burned (Fig. 1E), but there was no trend of increasing dominance over time, whereas both wildfire area and the number of houses in wildfires increased markedly from 1990 to 2020 (Fig. 1F).

Vegetation type affects fire intensity and behavior, and we asked whether vegetation type also affects the rate at which houses are destroyed. Of the 151,725 buildings (houses and other structures) that were exposed to wildfires from 2000 to 2013 (Fig. 1H), 11.3% were destroyed. However, buildings in evergreen and in mixed forests were almost twice as likely to be destroyed (20.1 and 22.9%, respectively). By contrast, the destruction rate for shrublands was similar to the average (12.7%), and rates for grasslands and deciduous forests were considerably lower (8.0 and 3.3%, respectively). Accordingly, logistic regression results showed that the proportion of these vegetation types within 1 km of buildings were all significant predictors of building loss ($p < 0.0001$; table S1). However, because the total area of grassland and shrubland fires is much larger than that of forest

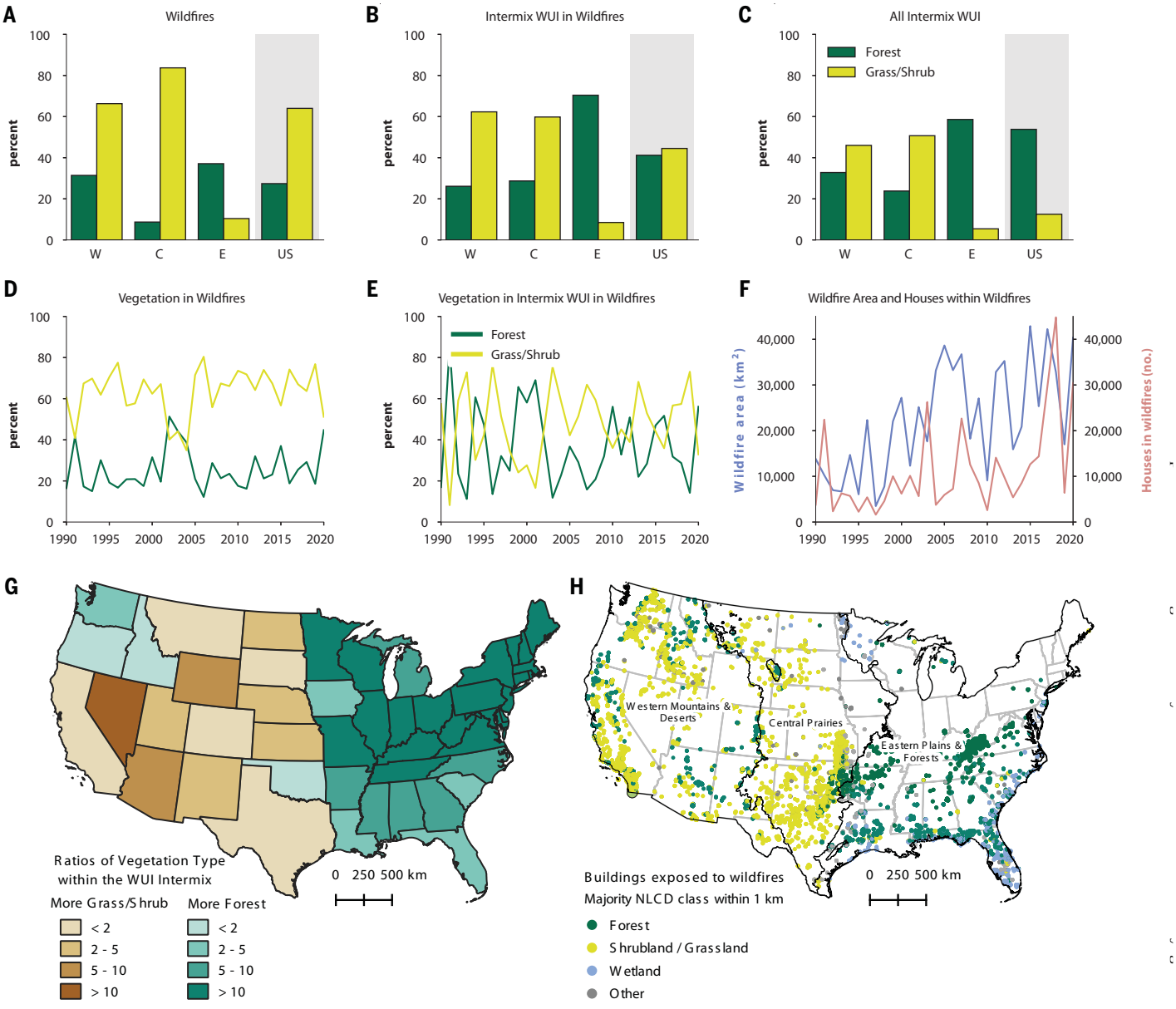


Fig. 1. The main vegetation types within wildfires, within the WUI, and near houses destroyed by wildfires. (**A** to **C**) The average percentage of forest, grassland, and shrubland land cover within the 1990–2020 wildfire perimeters in the western (W), central (C), eastern (E), and conterminous US (US, highlighted with gray shading) (A), within the portion of those wildfire perimeters that is 2020 intermix WUI (B), and within all (burned and unburned) 2020 intermix WUI (C). (**D** to **F**) The percentage of forest versus grassland and shrubland land cover within the entire 1990–2020 wildfire perimeters in the conterminous US over time (D) and within the portion of those wildfires that occurred in 2020 intermix WUI (E), and the annual wildfire area and the number of houses within wildfire perimeters over time (F). (**G** and **H**) The ratio of forest versus grassland and shrubland in the 2020 intermix WUI in each state (G) and the primary vegetation type surrounding buildings that were exposed to wildfires plus the boundaries of the three major ecoregions (H). Grasslands and shrublands dominate the intermix WUI in the Plains and the West, whereas forests dominate WUI in the East; however, because intermix WUI is far more widespread in the East, most intermix WUI in the US is dominated by forest, but intermix WUI within wildfires is dominated by grasslands and shrublands. Both burned area and the number of houses within wildfire perimeters have increased markedly since 1990, albeit with strong fluctuations from year to year. Most of the buildings that were destroyed by wildfires were also within grasslands and shrublands, especially in Texas, Oklahoma, and states of the West Coast. Data are provided in tables S2 and S3. NLCD, National Land Cover Database.

fires, 63.7% of houses destroyed in wildfires (10,890) were located in grasslands and shrublands, compared with only 33.1% in forests (5655). This pattern was most pronounced in the western US, which encompassed 69% of all the buildings destroyed by wildfires. There, 79.5% of all destroyed buildings were lost in grassland and shrubland fires. In the East, by contrast, 82.1% of destroyed buildings were lost in forest fires. In the West, even though forests had a high destruction rate (21.3%), only 2367 buildings were destroyed by forest fires compared with 9402 in grassland and shrubland fires. On average, a wildfire in an evergreen or mixed forest poses a considerably larger threat to an individual home than a wildfire in other vegetation types. We suggest that this is likely due to higher fire intensity but cannot rule out that houses near forests are built differently or that weather patterns associated with forest fires are inherently different. However, grassland

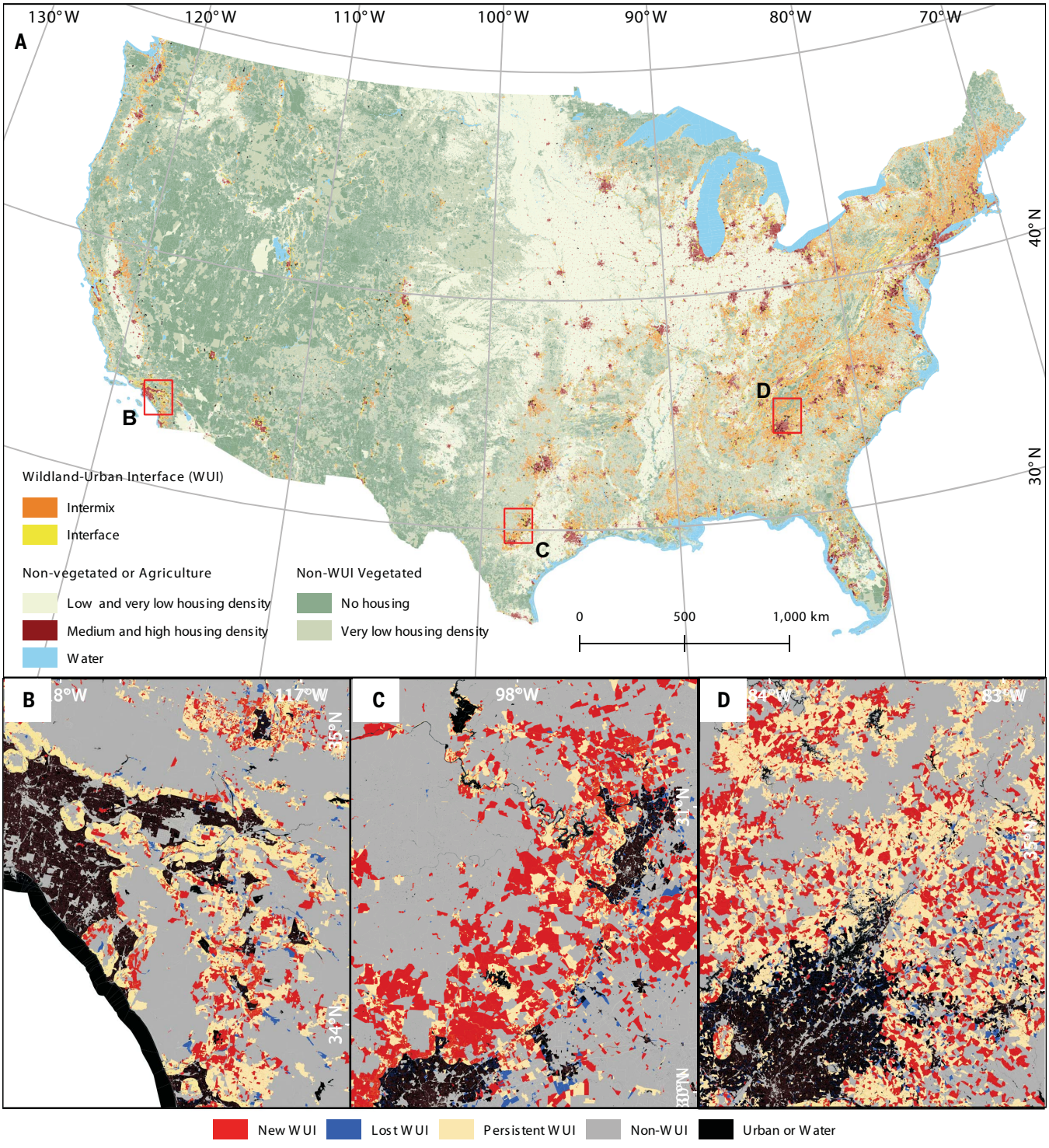


Fig. 2. The 2020 WUI and WUI change. (**A** to **D**) The WUI in 2020 across the conterminous US (A) and changes in WUI area from 1990 to 2020 near Los Angeles (B), San Antonio (C), and Atlanta (D).

and shrubland fires destroy far more homes overall because these fires are more widespread.

More houses within wildfire perimeters, both before and after fires

Over time, the number of houses within wildfire perimeters in the conterminous US more than doubled from 64,000 houses within 1990s fire perimeters, to 109,000 in the 2000s, and then to 148,000 in the 2010s. Of the houses within wildfire perimeters, three-quarters were within either the intermix or the interface WUI (74.9%), and that proportion was stable (75.2, 76.4, and 73.1% in the 1990s, 2000s, and 2010s, respectively), which highlights why the WUI is a focal area for wildfire concerns.

Increases in the number of houses exposed to wildfires can stem from either housing growth or increasing burned area. One indication of the strong effect of housing growth is that 39,000 of the 148,000 houses within 2010s

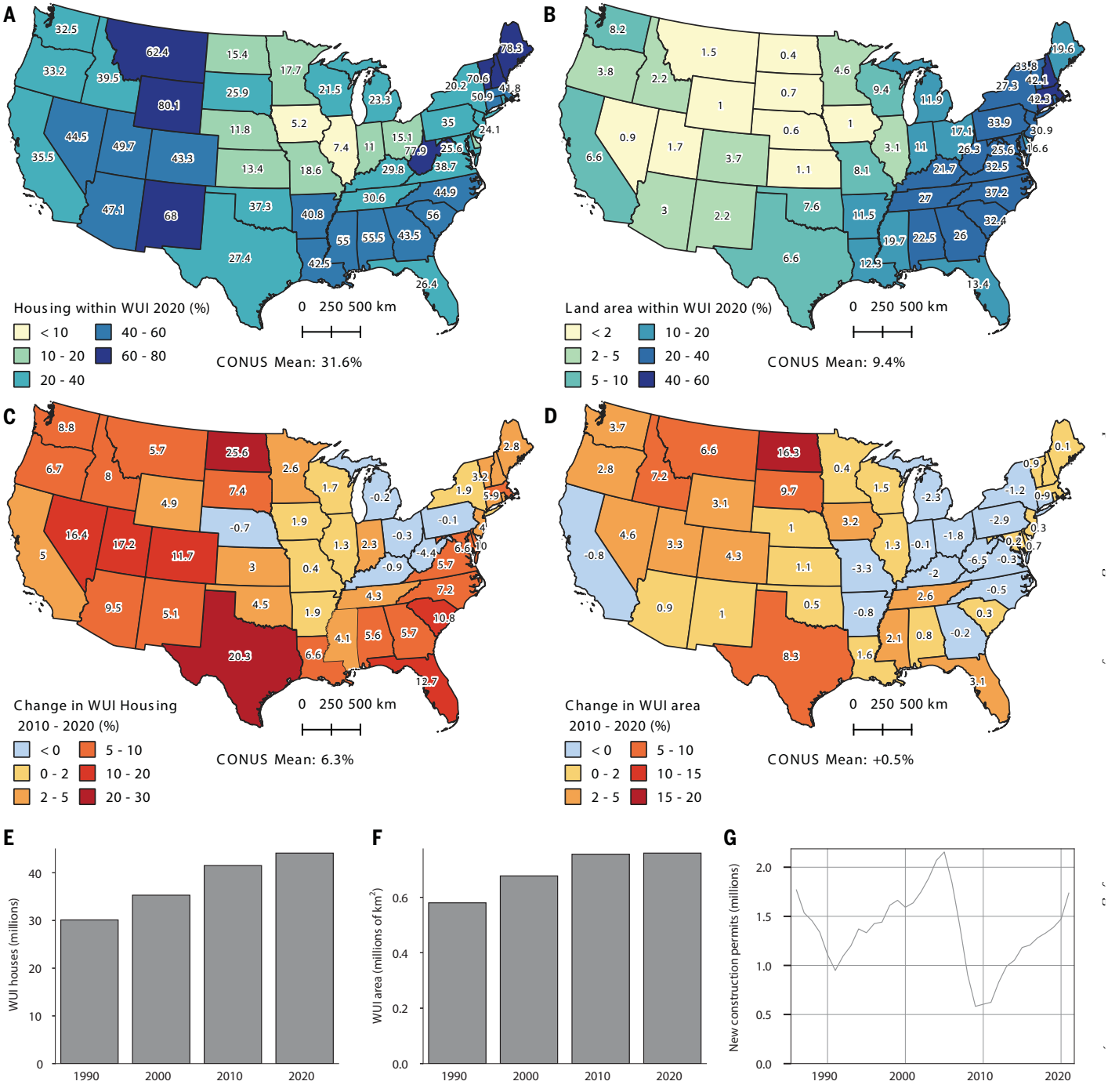


Fig. 3. The proportion of WUI and WUI change in each state and in each decade. (**A**) Proportion of houses within the WUI. (**B**) Proportion of the land area that is WUI. (**C** and **D**) Percentage growth of WUI houses (C) and WUI area from 2010 to 2020 (D). (**E**) Total numbers of WUI houses in 1990, 2000, 2010, and 2020. (**F**) WUI area in 1990, 2000, 2010, and 2020. (**G**) Number of new construction permits from 1990 to 2021. In many states, especially in the interior West and the Southeast, more than 40% of all houses are in the WUI. In the East, WUI also covers large proportions of most states. Growth in the number of WUI houses was highest in Texas and South Dakota and overall greater than growth in the amount of WUI area. WUI growth slowed after 2010, as indicated by the smaller increases in both WUI houses and WUI area from 2010 to 2020 than in prior decades. New construction permits peaked before the 2008 recession at about 2 million per year in 2004 to 2006, dropped to about 600,000 per year in 2009 and 2010, and recovered to 1.7 million in 2021. Data are provided in tables S4 and S5. CONUS, conterminous US.

wildfire perimeters were built after 1990, which is equivalent to 36% growth. However, burned area grew even more rapidly, from 114,000 to 274,000 km^2 during the 1990s and 2010s, respectively, which is equivalent to 240% growth, although most of this burned area contained few or no houses. Of the 84,000 additional houses located within 2010s fire perimeters compared with 1990s perimeters, we estimated that 47% were related to housing growth, a percentage that represents the proportion of the houses present in 2010 that were built after 1990 (see supplementary materials), and that the remaining 53% were related to increases in burned area. Overall, both housing growth

and burned area increases contributed substantially to the rise in the number of houses exposed to wildfires.

The increasing number of houses exposed to wildfires has not deterred new construction in the areas that burned. Indeed, past wildfire areas had higher housing growth after wildfires than nonfire areas (120 and 33% housing growth, respectively, within the 1990s and 2000s wildfire perimeters compared with 37 and 21% growth for the conterminous US as a whole). Within 1990s burn perimeters, there were only 64,000 houses in 1990 before the fires, but by 2020, there were 142,000 houses within those same perimeters, suggesting that those past fires were not a major deterrent to new development. Similarly, within the 2000s burn perimeters, there were 109,000 houses in 2000 but 145,000 by 2020. These numbers suggest that wildfire occurrence is not discouraging development. Even if another fire is unlikely in the immediate aftermath of a wildfire because of a lack of fuel, past wildfire areas can burn again within years or decades, depending on the vegetation type.

Growth of the WUI at-large

Given strong increases in both burned area and the number of houses exposed to wildfires and given that three-quarters of all houses within wildfire perimeters are in the intermix WUI, we asked how much the WUI at-large has grown. From 1990 to 2020, the US total WUI (both interface and intermix) increased by 31% in area, and the number of total WUI houses increased by 46%, from 30 million to 44 million (Fig. 2). Across the conterminous US, there were 2.6 million more houses in the total WUI in 2020 than in 2010, and in California alone, there were 244,000 more. Furthermore, WUI growth was concentrated in the southern US. Texas added 534,000 total WUI houses during the 2010s, which is considerably more than the 336,000 added during the 1990s. As a result, by 2020, there were 44.1 million, 5.1 million, and 3.2 million houses in the total WUI in the conterminous US, California, and Texas, respectively.

However, total WUI increased more slowly in terms of both houses and area during the 2010s than in the two prior decades (Fig. 3, A to D, and figs. S1 to S5). The absolute number of additional WUI houses during the 2020s was lower by more than half (2.6 million from 2010 to 2020 versus 5.7 million per decade from 1990 to 2010; Fig. 1E), and the rate of total WUI housing growth decreased by almost two-thirds (6.3 versus 17.4%). Furthermore, there was only a very limited increase in total WUI area during the 2010s (3900 km^2; Fig. 1F) compared with 1990 to 2010 (88,000 km^2 per decade) because almost all new WUI houses were built in areas that were already WUI in 2010.

Concomitant with slower total WUI growth, the total WUI became less of a focal area for housing growth during the 2010s. The percentage of all new homes that were built in the WUI (29.9%) was lower than that in the prior two decades (38.9%). Similarly, in California, the percentage of new houses that were built in the total WUI dropped to 34.3% during the 2010s from an average of 48.3% during the prior two decades. However, this reversal was not uniform. In Texas, housing growth in the total WUI was almost double the growth in non-WUI (20.3 versus 10.6%) during the 2010s, and 93.3% of all new houses in Texas during this most recent decade were built in WUI.

One reason why total WUI growth slowed was that the 2008 housing market collapse in the US greatly limited how many houses were constructed in the subsequent decade. From a high of 2.2 million in 2005, the number of permits for new construction plummeted to 905,000 in 2008 and 583,000 in 2009, a 73% decline (Fig. 3G). However, permits for new construction have rebounded because of strong demand and high housing prices, and new construction permits are approaching their pre-2008 peak (1.7 million in 2021, or 81% of peak). Regionally, the South recovered fastest (88% of peak), which contributed to rapid total WUI growth there, and the Midwest slowest (64%). Two-thirds of all new permits are still for single-family housing units (64.2%), but that percentage is lower than it was before the recession (78% in 2005), and in the Northeast, single-family units dropped to 39.5% of new housing construction. A shift toward more multiunit housing may be one reason for slower growth in WUI area. COVID-19 may have also influenced the number of WUI homes. The US Census data that we analyzed do not capture COVID-19–related changes in housing patterns, owing to the 1 April 2020 enumeration date, but other evidence suggests that there was a substantial migration toward rural areas (*22*).

Discussion

Our results highlight that homeowners and communities in the WUI have experienced wildfires more frequently in recent years. In many WUI areas, it is not a question of whether a wildfire will occur, but when. Proactive efforts to limit the loss of lives and houses during fires are thus of paramount importance (*23*). Hardening homes, so that fewer are destroyed during a fire, is one option. Overall, only 11.3% of houses within fire perimeters burn, and the fact that most houses survive suggests that it may be possible to reduce the rate at which houses are lost even further. Building materials, landscaping, and landscape position of houses all affect which houses burn (*24*). Limiting ignitions, especially during high-wind events when fires can spread quickly, is also important, and this requires public education [e.g., Smokey Bear, (*25*)], improvements to the electrical grid, and occasional shutdowns. When fires do occur, predicting their likely direction is critical for alerting residents rapidly and managing evacuations. All of these efforts are costly, and they would need to be done continuously; however, they are proven solutions to limiting the effects of fires.

One surprising result was that grassland and shrubland fires destroyed far more houses than forest fires, mainly because so much of the WUI in the West is dominated by grasslands and shrublands. The type of vegetation in the intermix WUI determines how quickly an area can reburn, which risk-management strategies are most appropriate, and how frequently they must be applied to reduce risk (*21*). In grasslands and shrublands, very different management approaches are needed to reduce risk in the intermix WUI than in forests (*26*). Prescribed fire can be an effective fuel treatment in grasslands and shrublands; however, fire needs to be applied frequently because fuel loads recover quickly (*27*, *28*). Further, fuel reduction and prescribed fires may not be advisable in all grasslands and shrublands (*29*), especially where fire-prone invasive species have replaced native vegetation, resulting in a positive-feedback loop and increased fire frequency (*30*–*32*). In some forests, thinning to reduce fuel loads, followed by prescribed surface fires, can reduce the likelihood of crown fires and revert the negative effects of past fire suppression (*33*). However, this strategy is only suitable for those forest types that are adapted to low-intensity surface fires, such as low-elevation forests in the mountainous West, where decades of fire suppression caused fuels to build up.

Irrespective of how much the total WUI grows in the future, the 44.1 million existing houses in the total WUI in 2020 mean that WUI wildfires will remain a major problem. Some future losses of houses due to wildfires are probably unavoidable and could be planned for by assisting affected homeowners and coordinating rebuilding efforts (*34*). If a policy goal is to limit risk to newly developed houses, this could be accomplished by both stricter building standards and land-use planning to avoid construction in the areas where fires are most likely to occur (*35*). California, for example, has comprehensive building requirements, but most states do not. For existing homes in the total WUI, risk could be reduced by hardening houses and other infrastructure (*36*) and by maintaining defensible space (*21*). Wildfire awareness of homeowners, as well as community involvement and actions, can be strengthened through outreach programs such as Firewise USA (*37*). Outreach, combined with enforcement, could also reduce unwanted ignitions, which cause most of the fires that lead to home losses (*15*, *38*). At the landscape scale, vegetation management, including fuel breaks, can reduce the potential for wildfires that have high intensity or that spread rapidly. Such efforts are most successful when facilitated

by agency-agency partnerships and multistakeholder collaboration (*39*).

Maintaining wildfires in ecosystems where they are a natural disturbance is an important management objective but challenging because they also pose major threats to many people, including health problems stemming from smoke, and their homes (*28*). Similarly, prescribed burning is an important tool for both fuel-load reductions and ecosystem restoration (*27*, *28*), and indigenous cultural burning fosters social restoration and helps maintain traditional knowledge and spiritual values (*40*); however, both are difficult to implement safely in WUI areas. Furthermore, although wildfires are the most pressing concern, there are many other environmental problems in the WUI. These include biodiversity loss due to habitat fragmentation, introductions of exotic species, noise and light pollution, and pets acting as subsidized predators (*41*, *42*). Similarly, the WUI is where the risk of transmission of zoonotic diseases, such as Lyme disease, is highest (*43*). WUI growth is hence a concern for a range of issues in addition to wildfires.

Our findings of high numbers of existing homes combined with rapid increases in burned areas, plus predictions for even more fires (*44*), mean that wildfire problems in the US will at least continue at present levels and likely increase. Similarly, increasing numbers of construction permits suggest that discussions about, and solutions to, wildfire risks to homes are urgently needed. Coordinated actions among homeowners, builders, communities, and public agencies alike, and a greater focus on wildfires in grasslands and shrublands, can limit loss of lives and houses.

REFERENCES AND NOTES

1. D. M. Bowman *et al*., *Science* **324**, 481–484 (2009).
2. V. C. Radeloff *et al*., *Ecol. Appl.* **15**, 799–805 (2005).
3. F. Schug *et al*., *Nature* **621**, 94–99 (2023).
4. National Interagency Fire Center, Statistics (2023); https://www.nifc.gov/fire-information/statistics.
5. J. T. Abatzoglou, A. P. Williams, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11770–11775 (2016).
6. T. Schoennagel *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4582–4590 (2017).
7. V. C. Radeloff *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3314–3319 (2018).
8. H. A. Kramer, M. H. Mockrin, P. M. Alexandre, V. C. Radeloff, *Int. J. Wildland Fire* **28**, 641–650 (2019).
9. E. Koo, P. J. Pagni, D. R. Weise, J. P. Woycheese, *Int. J. Wildland Fire* **19**, 818–843 (2010).
10. C. R. Evers, A. A. Ager, M. Nielsen-Pincus, P. Palaiologou, K. Bunzel, *Landsc. Urban Plan.* **182**, 55–66 (2019).
11. D. G. Brown, K. M. Johnson, T. R. Loveland, D. M. Theobald, *Ecol. Appl.* **15**, 1851–1863 (2005).
12. J. K. Balch, B. A. Bradley, C. M. D'Antonio, J. Gómez-Dans, *Glob. Change Biol.* **19**, 173–183 (2013).
13. R. K. Hagmann *et al*., *Ecol. Appl.* **31**, e02431 (2021).
14. J. K. Balch *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2946–2951 (2017).
15. A. D. Syphard, J. E. Keeley, A. H. Pfaff, K. Ferschweiler, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13750–13755 (2017).
16. J. S. Littell, D. L. Peterson, K. L. Riley, Y. Liu, C. H. Luce, *Glob. Change Biol.* **22**, 2353–2369 (2016).
17. A. L. Westerling, *Philos. Trans. R. Soc. Lond. Ser. B* **371**, 20150178 (2016).
18. S. Sachdeva, S. McCaffrey, *Int. J. Wildland Fire* **31**, 1089–1102 (2022).
19. S. L. Stephens, L. W. Ruth, *Ecol. Appl.* **15**, 532–542 (2005).
20. B. B. Hanberry, *Remote Sens.* **12**, 2966 (2020).
21. D. E. Calkin, J. D. Cohen, M. A. Finney, M. P. Thompson, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 746–751 (2014).
22. C. Dimke, M. C. Lee, J. Bayham, *West. Econ. Forum* **19**, 89–102 (2021).
23. M. A. Moritz *et al*., *Nature* **515**, 58–66 (2014).
24. A. D. Syphard, J. E. Keeley, A. Bar-Massada, T. J. Brennan, V. C. Radeloff, *PLOS ONE* **7**, e33954–e33954 (2012).
25. J. P. Prestemon, D. T. Butry, K. L. Abt, R. Sutphen, *For. Sci.* **56**, 181–192 (2010).
26. K. Wollstein, C. M. O'Connor, J. Gear, R. Hoagland, *Rangelands* **44**, 187–193 (2022).
27. P. M. Fernandes, *Curr. For. Rep.* **1**, 118–127 (2015).
28. P. M. Fernandes, H. S. Botelho, *Int. J. Wildland Fire* **12**, 117–128 (2003).
29. J. E. Keeley, C. J. Fotheringham, M. Morais, *Science* **284**, 1829–1832 (1999).
30. J. E. Keeley, P. van Mantgem, D. A. Falk, *Nat. Plants* **5**, 774–775 (2019).
31. A. D. Syphard, T. J. Brennan, J. E. Keeley, *Ecosphere* **10**, e02796 (2019).
32. M. L. Brooks *et al*., *Bioscience* **54**, 677–688 (2004).
33. A. A. Ager, C. R. Evers, M. A. Day, F. J. Alcasena, R. Houtman, *Landsc. Urban Plan.* **215**, 104212 (2021).
34. H. A. Kramer *et al*., *Land Use Policy* **107**, 105502 (2021).
35. R. K. Miller, C. B. Field, K. J. Mach, *Int. J. Disaster Risk Reduct.* **50**, 101686 (2020).
36. M. A. Moritz *et al*., *Front. For. Glob. Change* **5**, 848254 (2022).
37. National Fire Protection Association, Firewise USA: Residents reducing wildfire risk (2017); http://www.firewise.org/.
38. P. E. Higuera *et al*., *PNAS Nexus* **2**, pgad005 (2023).
39. L. Yung *et al*., *Fire Ecol.* **18**, 30 (2022).
40. J. W. Long, F. K. Lake, R. W. Goode, *For. Ecol. Manage.* **500**, 119597 (2021).
41. A. Bar-Massada, V. C. Radeloff, S. I. Stewart, *Bioscience* **64**, 429–437 (2014).
42. L. T. Kelly *et al*., *Science* **370**, eabb0355 (2020).
43. A. E. Larsen, A. J. MacDonald, A. J. Plantinga, *Am. J. Trop. Med. Hyg.* **91**, 747–755 (2014).
44. P. Gao *et al*., *Sci. Total Environ.* **789**, 147872 (2021).
45. V. C. Radeloff *et al*., The 1990-2020 wildland-urban interface of the conterminous United States – geospatial data. Third Edition. Forest Service Research Data Archive (2022); https://doi.org/10.2737/RDS-2015-0012-3.
46. H. A. Kramer, M. H. Mockrin, P. M. Alexandre, S. I. Stewart, V. C. Radeloff, Building loss and rebuilding within wildfire perimeters of the conterminous United States (2000-2013). Forest Service Research Data Archive (2023); https://doi.org/10.2737/RDS-2023-0040.

ACKNOWLEDGMENTS

Funding: This work was funded by US Department of Agriculture Forest Service grants 20-JV-11242308-089, 14-JV-11221636-094, 13-JV-11242309-050, and 11-JV-11221636-110; US Department of Agriculture McIntire Stennis award WIS03072; Joint Fire Science Program grant 14-JV-11221636-111; NASA LCLUC MuSLI grant 80NSSC21K0310; and US Geological Survey grant G20AC00010.
Author contributions: Conceptualization: V.C.R., D.H., A.M.P., M.H.M., T.J.H.; Methodology: V.C.R., D.H., T.J.H.; Investigation: D.H., A.C., P.M.A., H.A.K., V.C.R.; Visualization: V.C.R., F.S., S.M., D.H.; Funding acquisition: V.C.R., M.H.M., A.M.P., T.J.H.; Project administration: V.C.R., M.H.M.; Supervision: V.C.R., M.H.M.; Writing – original draft: V.C.R., A.M.P.; Writing – review and editing: All coauthors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The 1990–2020 WUI data are freely available via the Forest Service Data Research Archive (*45*). The data on building losses within wildfire perimeters are also freely available via the Forest Service Data Research Archive (*46*). Sources for other datasets are listed in the materials and methods. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ade9223
Materials and Methods
Figs. S1 to S6
Tables S1 to S5
References (*47*–*53*)

Submitted 19 September 2022; accepted 5 September 2023
10.1126/science.ade9223

PHOTONICS

Ultrafast mode-locked laser in nanophotonic lithium niobate

Qiushi Guo[1,2,3]*, Benjamin K. Gutierrez[4], Ryoto Sekine[1], Robert M. Gray[1], James A. Williams[1], Luis Ledezma[1,5], Luis Costa[1], Arkadev Roy[1], Selina Zhou[1], Mingchen Liu[1], Alireza Marandi[1,4]*

Mode-locked lasers (MLLs) generate ultrashort pulses with peak powers substantially exceeding their average powers. However, integrated MLLs that drive ultrafast nanophotonic circuits have remained elusive because of their typically low peak powers, lack of controllability, and challenges when integrating with nanophotonic platforms. In this work, we demonstrate an electrically pumped actively MLL in nanophotonic lithium niobate based on its hybrid integration with a III-V semiconductor optical amplifier. Our MLL generates ~4.8-ps optical pulses around 1065 nm at a repetition rate of ~10 GHz, with energies exceeding 2.6 pJ and peak powers beyond 0.5 W. The repetition rate and the carrier-envelope offset frequency of the output can be controlled in a wide range by using the driving frequency and the pump current, providing a route for fully stabilized on-chip frequency combs.

Mode-locked lasers (MLLs), which generate intense and coherent ultrashort optical pulses on picosecond and femtosecond timescales, have enabled numerous sciences and technologies in photonics such as extreme nonlinear optics (*1*), supercontinuum generation (*2*), optical atomic clocks (*3*), optical frequency combs (*4*), biological imaging (*5*), and photonic computing (*6*). Today's state-of-the-art MLLs are based on discrete fiber-based and free-space optical components and are expensive, power demanding, and bulky. Realizing MLLs on integrated photonic platforms promises widespread use of ultrafast photonic systems that are currently limited to table-top laboratory experiments. However, the performance of integrated MLLs has not been on par with their table-top counterparts, lacking the required peak intensities and degrees of controllability required for on-chip ultrafast optical systems (*7*). A major challenge lies in the simultaneous realization of large laser gain and an efficient mode-locking mechanism on integrated photonic platforms. Although III-V semiconductor gain media can be electrically pumped and exhibit a very high gain per unit length and high saturation powers (*8*), the conventional method of achieving mode-locking and short pulse generation on the same semiconductor chip requires a narrow range of pumping current, thus substantially limiting the output power and the tunability of the integrated MLLs (*9*, *10*).

To realize high-peak-power integrated MLLs, a promising approach consists of the hybrid integration of a semiconductor gain medium and an external mode-locking element based on electrooptic (EO) or nonlinear optical effects. Thin-film lithium niobate (TFLN) has emerged as a promising integrated nonlinear photonic platform with access to power-efficient and high-speed EO modulation (*11*, *12*) and strong quadratic [$\chi^{(2)}$] optical nonlinearity (*13*, *14*).

[1]Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Photonics Initiative, Advanced Science Research Center, City University of New York, NY, USA. [3]Physics Program, Graduate Center, City University of New York, New York, NY, USA. [4]Department of Applied Physics, California Institute of Technology, Pasadena, CA, USA. [5]Jet Propulsion Laboratory, Pasadena, CA, USA.
*Corresponding author. Email: qguo@gc.cuny.edu (Q.G.); marandi@caltech.edu (A.M.)

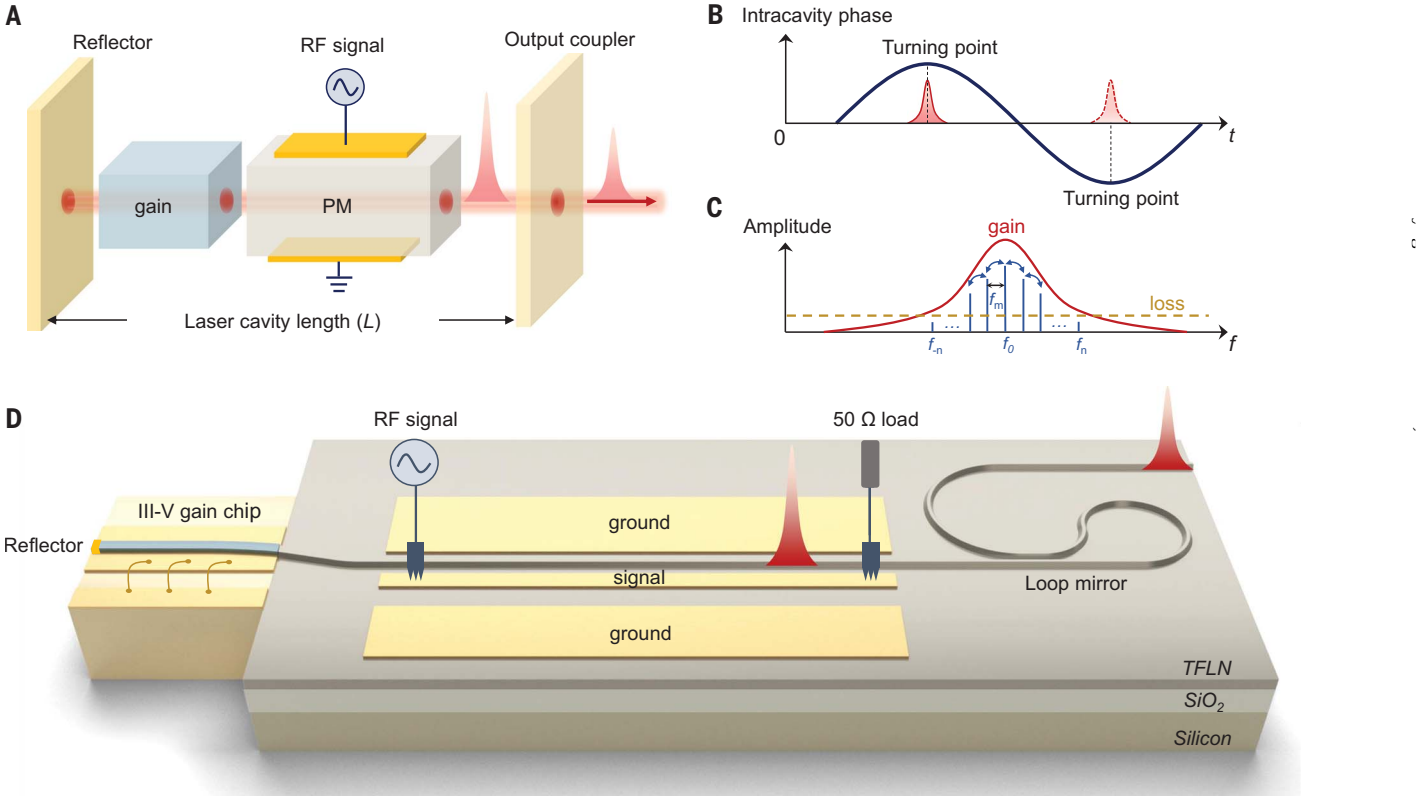


Fig. 1. Principle and design of integrated actively MLL laser. (**A**) Diagram of active mode-locking through intracavity phase modulation. (**B**) Illustration of mode-locking in the time domain. (**C**) Illustration of mode-locking in the frequency domain. (**D**) Schematic of the integrated actively MLL.

Hybrid integration of semiconductor gain with TFLN enables a strong interplay between the laser gain and the EO or nonlinear effects to achieve active or passive mode-locking with high efficiency and tunability. Moreover, many of the nonlinear and ultrafast optical functionalities such as supercontinuum generation (*15*), optical parametric oscillation (*16*–*18*), pulse shortening (*19*), all-optical switching (*20*), and quantum squeezing (*21*) can be realized in quasi–phase-matched LN nanophotonic devices with orders-of-magnitude lower peak powers compared with other platforms. Therefore, developing high-peak-power MLLs integrated into nanophotonic LN can enable a suite of nonlinear and ultrafast optical phenomena on a chip, promising integrated photonic systems with unprecedented performance and functionalities. In this work, we demonstrate a high-peak-power, electrically pumped, integrated, actively MLL by hybrid integration of III-V semiconductors and LN nanophotonics. Our MLL exploits the high laser gain of III-V semiconductors and the efficient active optical phase modulation in LN nanophotonic waveguides as the mode-locking mechanism. Such a design eliminates the complexities associated with realizing gain and saturable absorption on the same semiconductor chip, allowing a much higher output power and a wider tunability of the laser.

Mode-locked laser operating principle and design

Figure 1A shows the concept of active mode-locking by intracavity EO phase modulation. In the time domain, when a phase modulator (PM) is driven by a sinusoidal radio frequency (RF) signal at a frequency f_m, the intracavity phase modulation is equivalent to the cavity length modulation. Therefore, the laser cavity can be considered as having a moving end mirror with a sinusoidal motion at frequency f_m. When an optical signal inside the cavity strikes this moving end mirror and gets reflected back, its optical frequency acquires a Doppler shift. After successive round trips, these Doppler shifts will accumulate, resulting in no steady-state solution. However, when a short circulating pulse strikes the end mirror at either of the "turning points" where the mirror reverses its direction (the extremum of the phase variation as shown in Fig. 1B), it will not acquire a Doppler frequency shift, but instead, a small quadratic phase modulation or chirp (*22*). Thus, a steady-state optical pulse can be maintained in the laser cavity after successive round trips. Although in principle, optical pulses can occur at either of the two phase modulation extrema and acquire chirps of different signs, the dispersion in the cavity can compensate for the chirp imposed by the PM at one extremum, and further chirp the pulse formed at another extremum (*23*). The mode-locking mechanism can also be understood in the frequency domain (Fig. 1C). When the intracavity phase modulation frequency f_m matches the cavity free spectral range (FSR), the sidebands produced by each of the running axial modes are injected into the adjacent axial modes, resulting in the phase locking of adjacent modes. Notably, in MLLs, these modes will lase because of the presence of laser gain within the cavity, whereas in EO comb sources, they are generated by dispersing the energy from a single pump laser line (*24*–*27*).

Figure 1D shows the design of our integrated MLL based on this principle. In our MLL, an electrically pumped single-angled facet (SAF) GaAs gain chip is butt-coupled to a TFLN chip that contains an integrated EO PM and a broadband loop mirror. A Fabry-Perot laser cavity configuration is formed between the reflective facet on the left end of the SAF gain chip and the broadband loop mirror on the TFLN chip. Here, an integrated PM is preferred over a Mach Zehnder interferometer (MZI)–

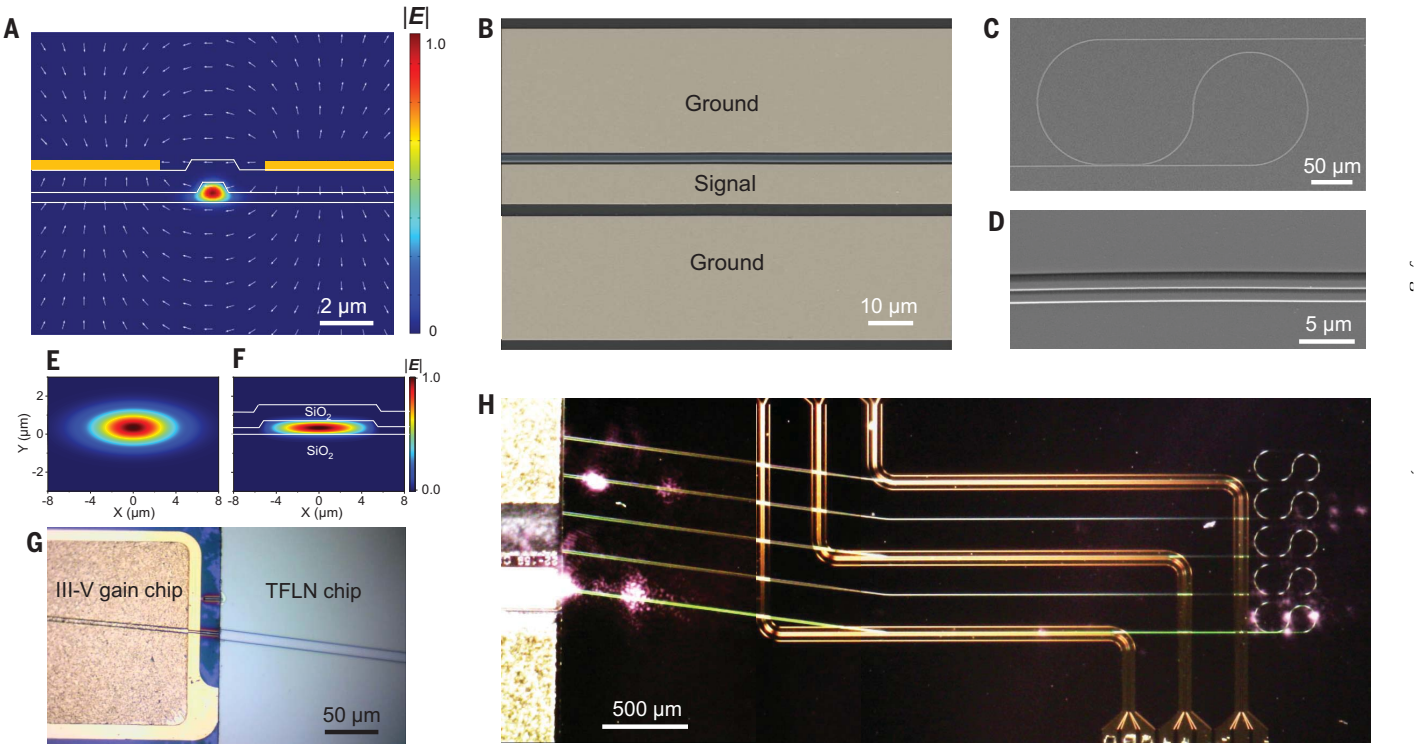


Fig. 2. Integrated actively MLL laser on TFLN. (**A**) Cross-sectional view of the PM region and the distribution of microwave field (white arrows) at 10 GHz and the optical field of the fundamental TE mode (color map) at 1065 nm. The RF electrodes are marked in yellow. |*E*|, normalized electric field. (**B**) False-colored SEM image of the PM region in the fabricated device. The RF electrodes are marked in yellow and the optical waveguide is marked in blue. (**C**) SEM image of the broadband loop mirror. (**D**) The curved coupling region of the broadband loop mirror. (**E** and **F**) The fundamental TE mode profile at 1065 nm in both the (E) SAF gain chip waveguide and (F) TFLN waveguide taper. (**G**) Optical microscope image showing the coupling region between the two chips. (**H**) Dark-field optical microscope image of the integrated MLL when operating.

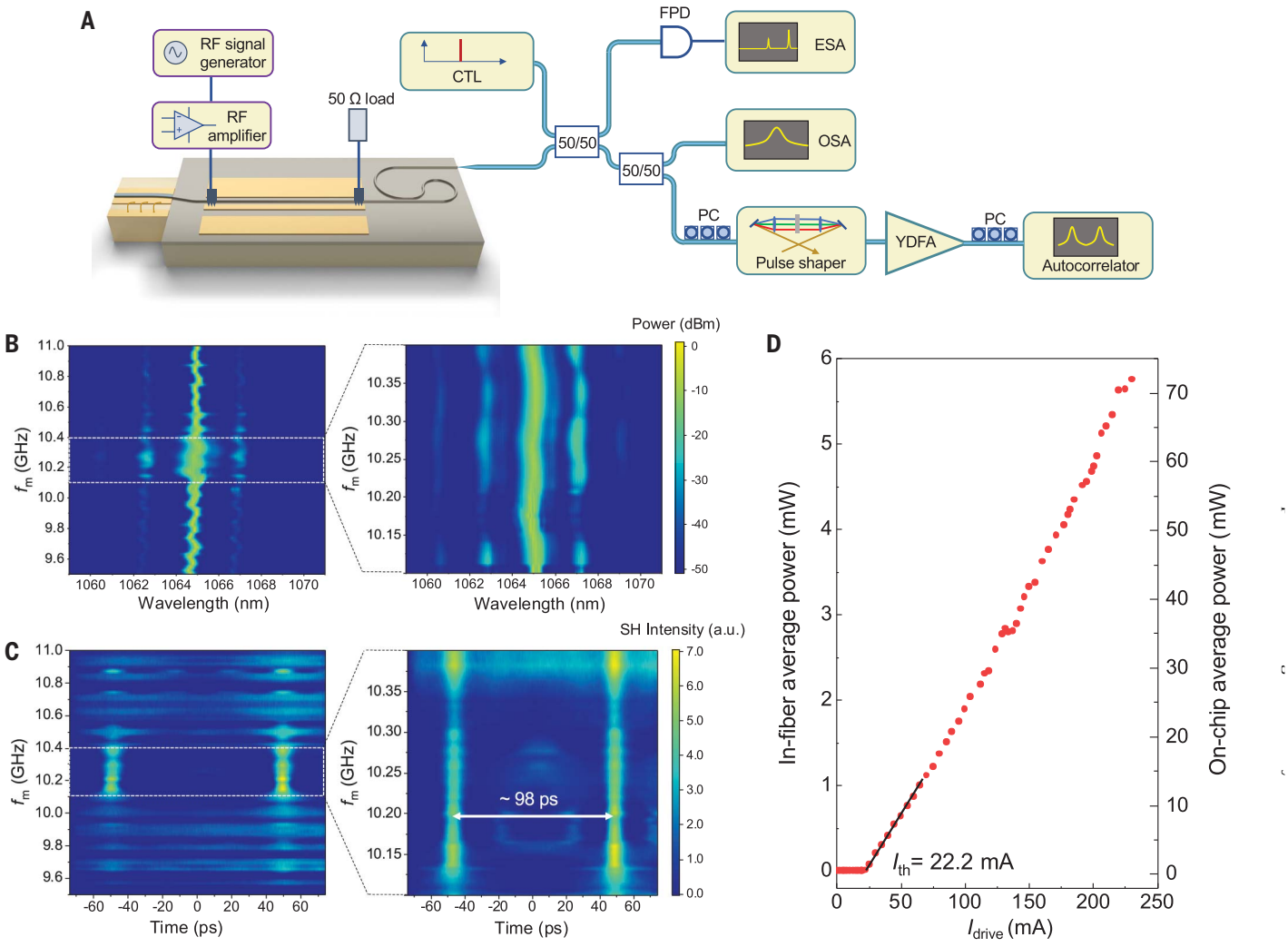


Fig. 3. Characterization of integrated actively MLL. (**A**) Schematic of the experimental setup. CTL, continuously tunable CW laser; FPD, fast photodetector; OSA, optical spectrum analyzer; ESA, electrical spectrum analyzer; YDFA, ytterbium-doped fiber amplifier. (**B**) The optical spectrum of the MLL output as a function of the RF driving frequency f_m. (**C**) Intensity autocorrelation of the MLL output as a function of the f_m. (**D**) Laser average output power versus I_{drive} when f_m = 10.17 GHz. I_{th}, threshold driving current.

based intensity modulator (IM) because the PM offers a lower insertion loss and avoids effects from the dc bias drift of the MZI modulator (*28*). For the PM, we used an RF coplanar waveguide (CPW) design with ground-signal-ground (GSG) configuration to ensure the transmission of the RF wave with minimal radiative loss.

Characterization of mode-locked lasers

We fabricated our devices on a 700-nm-thick X-cut magnesium oxide (MgO)–doped TFLN on a SiO_2/Si substrate. In the PM region (Fig. 2A), the RF CPW was fabricated on top of the SiO_2 cladding layer. Such a design allowed us to achieve high modulation efficiency (simulated $V_\pi L$ = 1.1 V·cm) by having a small gap (4 μm here) between the ground and signal electrodes and a significant overlap between the RF field and the optical field in the waveguide (*12*, *29*). We designed the geometry of the RF CPW to ensure a 50-Ω impedance around 10 GHz. Figure 2, B to D, shows the scanning electron microscope (SEM) images of the fabricated PM and the loop mirror. We adopted a curved coupling-region design in the loop mirror to ensure broadband reflection (see supplementary materials section I for details). Based on the length (1.5 mm) and the refractive index of the SAF gain chip around 1065 nm, we estimate that a ~3-mm-long TFLN waveguide including the loop mirror section can lead to a laser cavity FSR of ~10 GHz.

Figure 2E shows the 1065-nm fundamental transverse electric (TE)–mode profile in the waveguide of the SAF gain chip. To minimize the coupling loss between the SAF gain chip and the TFLN chip, the top width of the input facet of the TFLN waveguide was tapered out to be 10.3 μm. The 1065-nm fundamental TE-mode profile in the tapered TFLN waveguide is shown in Fig. 2F. This design ensures a maximal overlap with the optical mode produced by the SAF gain chip. We estimate a chip coupling loss of 3.4 to 3.9 dB in our experiments (see supplementary materials section II for details). The chip coupling loss can be further reduced by using a mode-size converter based on an inverted taper edge coupler embedded in a polymer waveguide (*30*, *31*). Figure 2G shows the microscope image of the chip coupling region after the alignment. When the SAF gain chip is electrically pumped with a driving current (I_{drive}) of 160 mA, we observed green light (the second harmonic of the 1065-nm light) inside the laser cavity (Fig. 1H), which indicated a high intracavity power around 1065 nm and a good alignment between the two chips.

We characterized the integrated actively MLL using the optical setup shown in Fig. 3A. We applied a ~280-mW sinusoidal RF signal to the left end of the CPW of the PM using the RF probe. The right end of the CPW is terminated

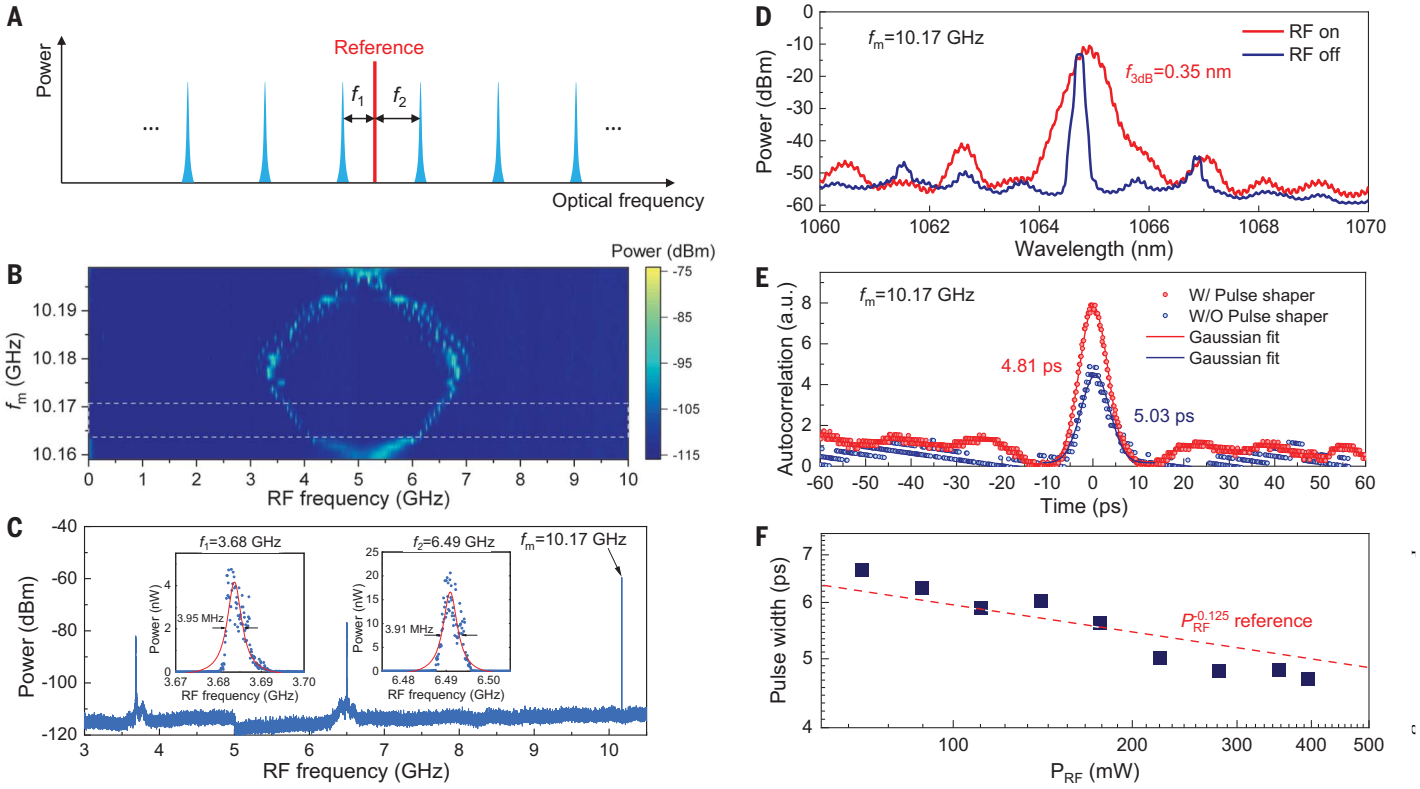


Fig. 4. Finding the mode-locking regime of integrated MLL. (**A**) Illustration of the generation of heterodyne beat notes. (**B**) Evolution of heterodyne beat notes with the f_m. The mode-locking regime is marked by the white dashed box. (**C**) Heterodyne beat notes measured at f_m = 10.17 GHz. (Insets) Zoomed-in view of the two beat notes at f_1 = 3.68 GHz and f_2 = 6.49 GHz. Blue symbols represent measured data and solid red curves indicate Lorentz fits. (**D**) Output optical spectra of the MLL when the RF drive at 10.17 GHz is on (red) and off (blue). The side lobes around 1062.7 and 1067 nm can be due to the reflections in the gap between the gain chip and the TFLN chip. The fast modulation can be due to the reflections within the gain chip. (**E**) Intensity autocorrelation traces of the MLL output measured at f_m = 10.17 GHz with (red) and without (blue) the external pulse shaper. Symbols represent measurement data and solid curves indicate Gaussian fits. a.u., arbitrary units. (**F**) Dependence of pulse width on P_{RF}. The red dashed line represents the $1/\sqrt[8]{P_{RF}}$ scaling law according to the HME.

by another RF probe with a commercially available 50-Ω load resistor. While the gain chip was pumped with an I_{drive} of 185.2 mA, we simultaneously collected the laser output spectra, the intensity autocorrelation of the laser output, and the heterodyne beat notes between two neighboring laser emission lines and a narrow-linewidth (~10 kHz) reference continuous wave (CW) tunable laser. As shown in Fig. 3B, when we scanned the f_m, the laser output exhibited a clear spectral broadening between 10.1 and 10.4 GHz (labeled by the white dashed box). Within this f_m range, two distinct intensity autocorrelation peaks separated by ~98 ps emerged (Fig. 3C), indicating that optical pulses are formed. At an f_m of 10.17 GHz, we measured the laser output power from the output facet of the TFLN chip. As shown in Fig. 3D, the laser exhibits a low threshold I_{drive} of 22 mA. Given the measured coupling loss of ~11 dB between the TFLN waveguide and the single-mode lensed fiber, the on-chip laser output average power is more than 50 mW when the I_{drive} is greater than 180 mA.

Heterodyne beat notes were used to characterize the mode-locking and resulting frequency comb. As illustrated in Fig. 4A, when the frequency of the reference CTL is resting in between the two neighboring comb lines of the MLL near the center of its spectrum, two RF beat notes at f_1 and f_2 are generated on the fast detector. As shown in Fig. 4B, when f_m is between 10.165 and 10.173 GHz, as labeled by the white dashed box, two spectrally narrow beat notes are observed. This suggests that within this range of f_m, the laser operates in the mode-locked regime, producing ultrashort optical pulses with high coherence. When f_m is slightly detuned between 10.165 and 10.173 GHz, f_1 and f_2 can shift significantly with f_m (Fig. 4B). This indicates that the carrier frequency of the MLL sensitively depends on f_m. However, when the f_m is further detuned from the cavity FSR, the MLL exhibits a transition to a turbulent regime (*32*), which is manifested by multiple noisy beat notes around f_1 and f_2 in Fig. 4B. In the turbulent regime, the laser can still emit ultrashort pulses as shown in Fig. 3C, albeit with low coherence.

As shown in Fig. 4C, at f_m = 10.17 GHz, we obtained two spectrally narrow RF beat notes at f_1 = 3.68 GHz and f_2 =6.49 GHz, with full width at half maximum (FWHM) linewidths of 3.95 and 3.91 MHz, respectively. Given that the RF drive has a very small phase noise, and no active locking of the laser cavity was used here, the linewidths of the heterodyne beat notes can be mainly limited by the drift of pulse carrier frequency. As shown in Fig. 4D, when a 280-mW RF drive at 10.17 GHz was applied to the PM, significant spectral broadening was observed. The pulse spectrum was centered at 1064.9 nm and the FWHM of the spectrum was 0.35 nm. Meanwhile, the intensity autocorrelation trace (Fig. 4E) indicated that the MLL produced one strong pulse at one of the modulation turning points, whereas the other pulse was significantly suppressed. We fit the intensity autocorrelation trace with a Gaussian function becasue active mode locking produces Gaussian pulses according to the Haus master equation (HME) (*33*). The fitting yielded a pulse width of 4.81 ps with the external pulse shaping, and 5.03 ps without. Because the pulse shaper can compensate for the chirp on the output pulse and the additional chirp imposed by the fibers and the YDFA, we expect the output pulse width directly after the

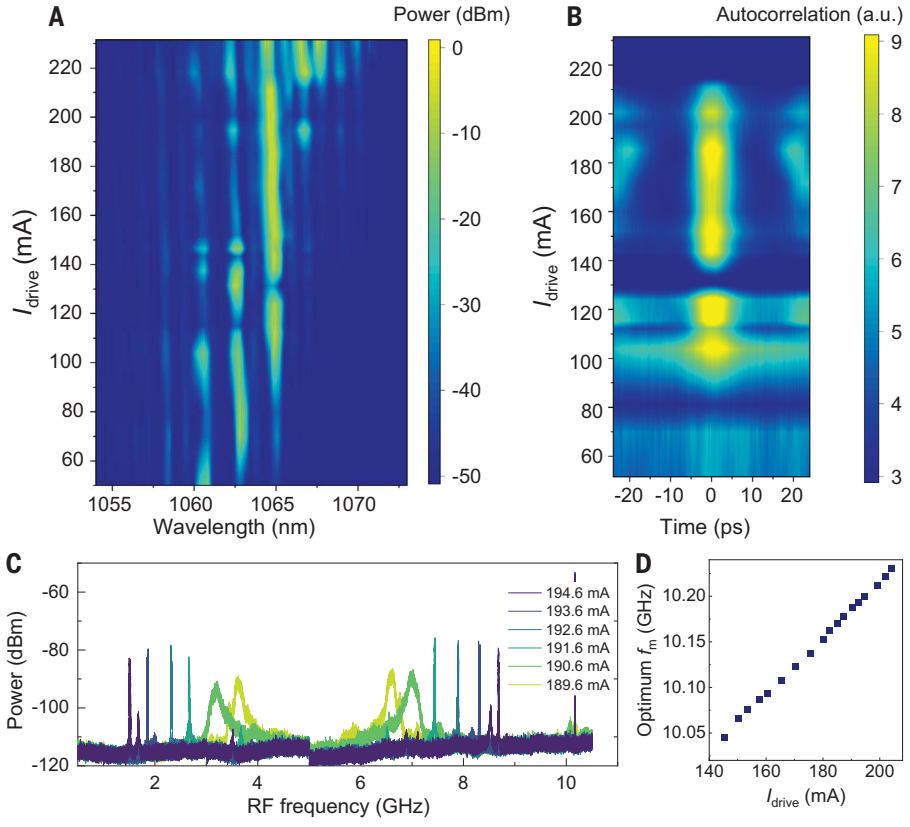


Fig. 5. Current tuning of integrated actively MLL. (**A**) The optical spectrum of the MLL output as a function of I_{drive}. (**B**) Autocorrelation trace of the MLL output as a function of I_{drive}. a.u., arbitrary units. (**C**) Tuning of the heterodyne beat notes by the I_{drive}. In (A) to (C), f_m is fixed at 10.18 GHz. (**D**) Dependence of optimum f_m for mode-locking on I_{drive}.

MLL facet to be between 4.81 ps and 5.03 ps. The pulse width of 4.81 ps after pulse shaping corresponds to a time-bandwidth product of 0.445, which is very close to the transform-limited time-bandwidth product (0.44) of a Gaussian pulse (*34*). To conservatively estimate the pulse energy and peak power, we used the measured output average power of 53 mW at I_{drive} = 185.2 mA and assumed that both pulses exist in the cavity. Hence, the output pulse energy of our MLL was at least 2.6 pJ and the pulse peak power was greater than 0.51 W.

We further studied the pulse width limits of our MLL. First, we found that the measured pulse only slightly decreased when the RF driving power P_{RF} was increased (Fig. 4F), which is in good agreement with the $1/\sqrt[3]{P_{RF}}$ scaling law (red dashed line) according to the HME (*35*). We also found that further increasing the RF power will not shorten the pulse significantly. Instead, it can lead to laser instability due to RF heating. The HME with cavity group velocity dispersion, neglecting nonlinear effects within the laser cavity, predicts a minimum pulse width of ~2.5 ps (*36*). The experimentally measured pulse width was wider, likely due to several factors that are not captured by the HME such as the complex carrier dynamics and two-stage gain recovery in the III-V gain medium (*37*) and gain bandwidth narrowing from intracavity etalon (*22*) or hole burning effects (*38*) (see supplementary materials section IV for detailed analysis).

Electrical tuning of mode-locked lasers

The I_{drive} can serve as an important tuning knob of our MLL. Figure 5, A and B, shows the dependence of the output spectra and autocorrelation of the MLL on the I_{drive} with 280-mW RF drive fixed at 10.17 GHz. Within a wide range of I_{drive} (140 to 205 mA), optical pulses can be formed inside the laser. In addition, the carrier frequency of the MLL blueshifted by ~0.3 nm as the I_{drive} was increased from 140 to 200 mA (Fig. 5A). This was likely caused by the blueshift of the peak wavelength of the gain spectrum owing to band filling and screening effects induced by carrier injection (*39*). We also investigated the effect of I_{drive} on the coherence property and the f_{rep} of the laser. We kept the RF drive fixed at 280 mW and 10.18 GHz. As the I_{drive} was tuned from 189.6 to 194.6 mA, the laser transitioned from the turbulent to the mode-locked regime, and then back to the turbulent regime (Fig. 5C). These results suggest that, with a frequency-stable reference CW laser and active feedback on I_{drive}, it may be possible to lock the carrier frequency of the MLL and operate the device as a stable frequency comb, as the repetition rate (f_{rep}) of the MLL has already been locked by the external RF oscillator. As shown in Fig. 5D, when we widely varied I_{drive} from 144 to 204 mA, the optimum f_m that enables mode-locking with high coherence could be varied from 10.04 to 10.23 GHz, indicating that the f_{rep} of the laser could also be adjusted by ~200 MHz. Moreover, the optimum f_m increased almost linearly with I_{drive}, which resulted from an increase of the cavity FSR caused by carrier injection in the gain medium. Although further increasing the I_{drive} would not lead to laser output power saturation, we did not observe pulse formation at higher I_{drive} beyond 205 mA. This is likely attributed to the significant detuning of the cavity FSR or the instability of laser mode-locking due to more pronounced self-phase modulation in the gain medium (*37*).

Conclusions and outlook

We have demonstrated an integrated actively MLL in nanophotonic lithium niobate operating around 1065 nm, which offers the highest output pulse energy and peak power of any integrated MLLs in nanophotonic platforms (fig. S9 and table S2). Our MLL allows for a wide tuning range of the laser f_{rep} of ~200 MHz and precise control of the laser's coherence properties. The current tuning capability of our MLL indicates that active feedback to I_{drive} can achieve simultaneous locking of the carrier frequency and f_{rep} of the MLL. This allows the MLL to operate as a stable frequency comb with locked carrier frequency offset (f_{CEO}) and f_{rep}. Finally, seamless integration of our high–peak power MLL with other $\chi^{(2)}$ nonlinear optical functionalities provided by TFLN offers opportunities for realizing new integrated photonic systems, such as fully integrated supercontinuum sources, self-referenced frequency combs, and atomic clocks.

REFERENCES AND NOTES

1. M. Wegener, *Extreme Nonlinear Optics: An Introduction* (Springer Science & Business Media, 2005).
2. J. M. Dudley, G. Genty, S. Coen, *Rev. Mod. Phys.* **78**, 1135–1184 (2006).
3. A. D. Ludlow, M. M. Boyd, J. Ye, E. Peik, P. O. Schmidt, *Rev. Mod. Phys.* **87**, 637–701 (2015).
4. T. Udem, R. Holzwarth, T. W. Hänsch, *Nature* **416**, 233–237 (2002).
5. D. W. Piston, *Trends Cell Biol.* **9**, 66–69 (1999).
6. A. Marandi, Z. Wang, K. Takata, R. L. Byer, Y. Yamamoto, *Nat. Photonics* **8**, 937–942 (2014).
7. M. L. Davenport, S. Liu, J. E. Bowers, *Photon. Res.* **6**, 468 (2018).
8. P. J. Delfyett *et al.*, *IEEE J. Quantum Electron.* **28**, 2203–2219 (1992).
9. A. Hermans *et al.*, *APL Photonics* **6**, 096102 (2021).
10. E. Vissers *et al.*, *Opt. Express* **30**, 42394–42405 (2022).
11. C. Wang *et al.*, *Nature* **562**, 101–104 (2018).
12. M. Xu *et al.*, *Optica* **9**, 61 (2022).
13. C. Wang *et al.*, *Optica* **5**, 1438 (2018).
14. J. Lu *et al.*, *Optica* **6**, 1455 (2019).
15. M. Jankowski *et al.*, *Optica* **7**, 40 (2020).

16. L. Ledezma *et al*., *Sci. Adv.* **9**, eadf9711 (2023).
17. J. Lu *et al*., *Optica* **8**, 539 (2021).
18. T. P. McKenna *et al*., *Nat. Commun.* **13**, 4532 (2022).
19. A. Roy *et al*., *Nat. Photonics* **16**, 162–168 (2022).
20. Q. Guo *et al*., *Nat. Photonics* **16**, 625–631 (2022).
21. R. Nehra *et al*., *Science* **377**, 1333–1337 (2022).
22. D. Kuizenga, A. Siegman, *IEEE J. Quantum Electron.* **6**, 694–708 (1970).
23. R. Nagar, D. Abraham, G. Eisenstein, *Opt. Lett.* **17**, 1119–1121 (1992).
24. M. Zhang *et al*., *Nature* **568**, 373–377 (2019).
25. M. Yu *et al*., *Nature* **612**, 252–258 (2022).
26. Y. Hu *et al*., *Nat. Photonics* **16**, 679–685 (2022).
27. A. Rueda, F. Sedlmeir, M. Kumari, G. Leuchs, H. G. L. Schwefel, *Nature* **568**, 378–381 (2019).
28. J. P. Salvestrini, L. Guilbert, M. Fontana, M. Abarkan, S. Gille, *J. Lightwave Technol.* **29**, 1522–1534 (2011).
29. M. Jin, J. Chen, Y. Sua, P. Kumar, Y. Huang, *Opt. Lett.* **46**, 1884–1887 (2021).
30. C. Hu *et al*., *Opt. Express* **29**, 5397–5406 (2021).
31. E. Vissers, S. Poelman, C. O. de Beeck, K. Van Gasse, B. Kuyken, *Opt. Express* **29**, 15013–15022 (2021).
32. F. Kärtner, D. Zumbühl, N. Matuschek, *Phys. Rev. Lett.* **82**, 4428–4431 (1999).
33. U. Keller, R. Paschotta, *Ultrafast Lasers* (Springer, 2021).
34. A. E. Siegman, *Lasers* (University Science Books, 1986).
35. H. A. Haus, *IEEE J. Sel. Top. Quantum Electron.* **6**, 1173–1185 (2000).
36. Q. Guo, data and computer code for "Ultrafast mode-locked laser in nanophotonic lithium niobate", version 2, Figshare (2023); https://doi.org/10.6084/m9.figshare.24208434.
37. G. P. Agrawal, N. A. Olsson, *IEEE J. Quantum Electron.* **25**, 2297–2306 (1989).
38. A. Uskov, J. Mork, J. Mark, *IEEE J. Quantum Electron.* **30**, 1769–1781 (1994).
39. S. Schmitt-Rink, D. S. Chemla, D. A. Miller, *Phys. Rev. B Condens. Matter* **32**, 6601–6609 (1985).

ACKNOWLEDGMENTS

The device nanofabrication was performed at the Kavli Nanoscience Institute (KNI) at Caltech. The authors thank K. Vahala for loaning equipment. Q.G. thanks M.Xu for the helpful discussions. **Funding:** The authors acknowledge support from ARO grant no. W911NF-23-1-0048, NSF grant nos. 1846273 and 1918549, AFOSR award FA9550-23-1-0755, and NASA JPL. The authors thank NTT Research for their financial and technical support. **Author contributions:** Q.G. and A.M. conceived the project. Q.G. fabricated the devices with assistance from R.S. Q.G. performed the measurements and numerical simulation. R.S., J.W., B.K.G., R.M.G., L.L., L.C., and S.Z. assisted with the measurements. B.K.G., R.M.G., A.R., and M.L. helped with the numerical simulation and data analysis. Q G. wrote the manuscript with input from all authors. A.M. supervised the project.

Competing interests: Q.G. and A.M. are inventors on a patent application (US patent application no. 17/500,425) that covers the concept and implementation of the actively MLL in this work. L.L. and A.M. are involved in developing photonic integrated nonlinear circuits at PINC Technologies Inc. L.L. and A.M. have an equity interest in PINC Technologies Inc. The remaining authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. The data files supporting the plots in the main text and the computer code for simulating the MLL are available at Figshare (*36*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adj5438
Materials and Methods
Supplementary Text
Figs. S1 to S9
Tables S1 and S2
References (*40*–*48*)

Submitted 2 July 2023; accepted 4 October 2023
10.1126/science.adj5438

ANTHROPOLOGY

Early *Homo erectus* lived at high altitudes and produced both Oldowan and Acheulean tools

Margherita Mussi[1,2,3]*, Matthew M. Skinner[4,5], Rita T. Melis[2,6], Joaquín Panera[2,7,8], Susana Rubio-Jara[2,7,8,9], Thomas W. Davies[4,10], Denis Geraads[11], Hervé Bocherens[12,13], Giuseppe Briatico[1,2,12], Adeline Le Cabec[14,10], Jean-Jacques Hublin[10,15], Agness Gidna[16], Raymonde Bonnefille[17], Luca Di Bianco[2], Eduardo Méndez-Quintas[2,8,18]

In Africa, the scarcity of hominin remains found in direct association with stone tools has hindered attempts to link *Homo habilis* and *Homo erectus* with particular lithic industries. The infant mandible discovered in level E at Garba IV (Melka Kunture) on the highlands of Ethiopia is critical to this issue because of its direct association with an Oldowan lithic industry. Here, we used synchrotron imaging to examine the internal morphology of the unerupted permanent dentition and confirmed its identification as *H. erectus*. Additionally, we used revised paleomagnetic ages to show that (i) the mandible in level E is ~2 million years old and represents one of the earliest *H. erectus* fossils and that (ii) overlying level D, ~1.95 million years old, contains the earliest known Acheulean assemblage.

The Melka Kunture complex consists of a cluster of prehistoric sites that extends over tens of square kilometers in the Upper Awash, on the Ethiopian plateau at ≥2000 m above sea level (asl) (Fig. 1A). At this high altitude, the Pleistocene climate would have been cooler and wetter than at the lower elevations of the Rift Valley, from where most fossil evidence for early species of *Homo* derives. The paleovegetation belonged to the Afromontane complex, ranging from forest to grassland and bushland (*1*), and the plant species were distinct from those of the African savanna, which developed at lower elevations in drier and warmer environments. The degree of topographic relief of the region meant that the highlands were relatively isolated, leading to the appearance of endemic mammal subspecies (*2*). In 1981, an infant mandible was discovered in level E of the Garba IV site, although which species of hominin it represents has been a subject of debate (*3*). In this study, we review the geochronological context of the Garba deposits, characterize the Oldowan and Acheulean lithic assemblages, present the paleoecological context of the site on the basis of analysis of the faunal assemblage and palynology, and reassess the taxonomic affinity of the mandible. Our aim is to determine when, how, and which hominin species was first able to adjust to the high-altitude environment of the highlands of Melka Kunture during the Early Pleistocene and the associated transition from the Oldowan to the Acheulean technocomplex. Broadening our knowledge of the environment at the time and not limiting it to the savanna and to lower altitudes is relevant to understanding how *H. erectus* was able to expand beyond Africa.

Stratigraphy and chronology

The sedimentary sequence of the multilayered Garba IV site records hominin activity on lag deposits close to shallow ponds and pools that are situated within a floodplain affected by the migrating paleo-Awash channel and associated flood events and also influenced by the deposition of volcanic ashes. The site is at the mouth of a small gully, the Garba gully, incised by a seasonal tributary of the Awash River (fig. S1). We refer here to three main archaeological levels at the base of the stratigraphic sequence of the gully (Fig. 1B and figs. S1 to S3): levels F and E with faunal and hominin remains and an Oldowan lithic industry, and level D with faunal remains and an Early Acheulean industry. $^{40}Ar/^{39}Ar$ and preliminary magnetostratigraphic analyses had previously dated these levels to shortly before and shortly after 1.7 million ± 0.2 million years ago (Ma) (*4*) but the age was not well constrained. A new magnetostratigraphic study has recently established the age of the levels (*6*). In this sequence, both the Olduvai/Matuyama boundary (1.925 Ma) and the Matuyama/Reunion boundary (2.116 Ma) are identified, which provide a maximum and a minimum age for the three archaeological layers. The Olduvai/Matuyama boundary was observed above level D. Using an estimated depositional rate and a model of sediment accumulation, the specific ages have been calculated as follows: level F, 2.02 (−0.095; +0.096) Ma; level E, 2.00 (−0.0075; +0.1) Ma; and level D, 1.95 (−0.025; +0.1) Ma (*6*) (Fig. 1, C and D). These much older dates for levels D to F have considerable implications for the interpretation of the archaeology and the hominin fossil found at the site.

Archaeology

Level E at Garba IV (now dated to ~2.00 Ma) was first tested in 1981, and 34 m^2 were eventually exposed (fig. S4 and figs. S8 to S10), with an additional exposure of 12 m^2 of underlying level F (now dated to ~2.02 Ma) (fig. S4 and figs. S8 to S10). Both yielded abundant Oldowan tools (*7*), and the associated mandible represents one of the only hominin remains directly associated with artefacts in the Early Pleistocene (table S1). It was excavated in level E, a sealed level within a well-established stratigraphic sequence (Fig. 1, C and D; fig. S2; and figs. S9 to S11). The artefacts were mostly produced on obsidian cobbles and pebbles preferentially selected from the rocks available in the local alluvial deposits (Fig. 2A and figs. S12 and S13). Many of these artefacts are made from small flakes, mostly 20 to 50 mm in length, with ~30% retouched and transformed into small, pointed tools (Fig. 2A, 3 to 5, and fig. S13).

In overlying level D, excavated since 1973 to some 75 m^2 and redated to ~1.95 Ma, the ~7000 lithic tools (table S1) differ from those of the underlying Oldowan assemblage owing to the production of large flakes (>100 mm in maximal length or width). There are true handaxes, cleavers, and other types of large cutting tools (LCTs), such as massive scrapers or knives shaped on large flakes (Fig. 2B, 6 to 12, and figs. S15 to S19). The large flakes and

[1]Dipartimento di Scienze dell'Antichità, Sapienza Università di Roma, 00185 Roma, Italy. [2]Italo-Spanish Archaeological Mission at Melka Kunture and Balchit, Italy-Spain. [3]ISMEO – The International Association for Mediterranean and Oriental Studies, 00186 Roma, Italy. [4]School of Anthropology and Conservation, University of Kent, Canterbury, CT2 7NR, UK. [5]Centre for the Exploration of the Deep Human Journey (CSDHJ), University of the Witwatersrand, Braamfontein 2000, Johannesburg, South Africa. [6]Dipartimento di Scienze Chimiche e Geologiche, Università di Cagliari, Cittadella Universitaria, 09042 Monserrato, Italy. [7]Departamento de Prehistoria, Historia Antigua y Arqueología, Facultad de Geografía e Historia, Universidad Complutense de Madrid, 28040 Madrid, Spain. [8]IDEA, Instituto de Evolución en África, Universidad de Alcalá de Henares, 28010 Madrid, Spain. [9]Centro Nacional de Investigación sobre la Evolución Humana, 09002 Burgos, Spain. [10]Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. [11]Centre de Recherche en Paléontologie – Paris, Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, 75231 Paris Cedex 05, France. [12]Department of Geosciences, Eberhard Karls University of Tübingen, 72074 Tübingen, Germany. [13]Senckenberg Centre for Human Evolution and Palaeoenvironment, 72076 Tübingen, Germany. [14]Univ. Bordeaux, CNRS, Ministère de la Culture, PACEA (UMR 5199), F-33600 Pessac, France. [15]Chaire de Paléoanthropologie, CIRB (UMR 7241–U1050), Collège de France, 75231 Paris Cedex 05, France. [16]Department of Cultural Heritage, Ngorongoro Conservation Area, Arusha, Tanzania. [17]CEREGE, Université Aix-Marseille, CNRS, IRD, INRAE, Technopole Arbois-Méditerranée, 13545 Aix-en-Provence Cedex 04, France. [18]GEAAT, Grupo de Estudos de Arqueoloxía, Antigüidade e Territorio. Universidade de Vigo, Facultade de Historia, Campus As Lagoas, 32004 Ourense, Spain.
*Corresponding author. Email: margherita.mussi@fondazione.uniroma1.it

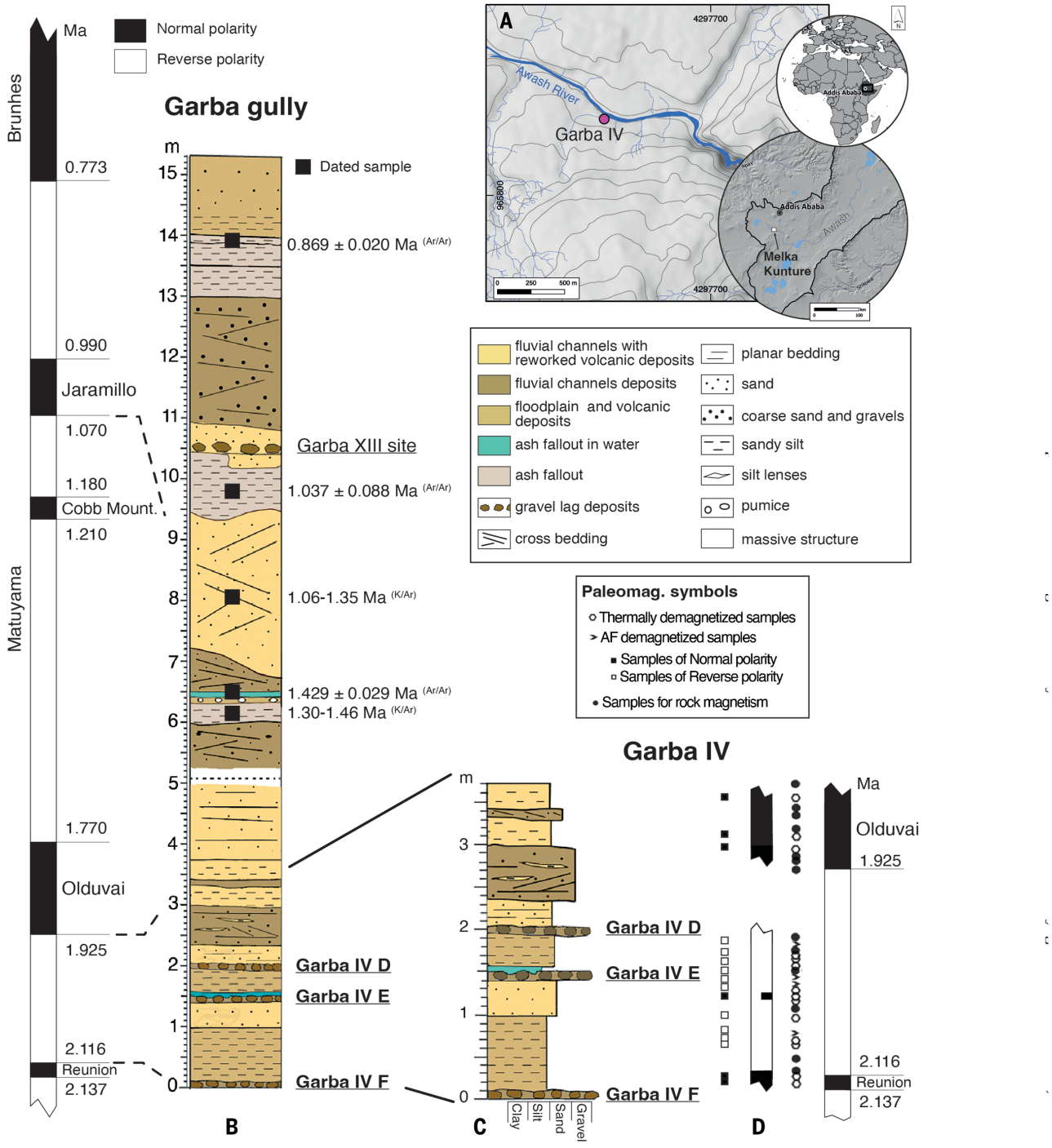


Fig. 1. Location, magnetostratigraphy, and chronostratigraphy of Garba IV. (**A**) Location map of the Garba IV site (Melka Kunture, Upper Awash, Ethiopia). (**B**) The stratigraphy of the Garba gully, with both Garba IV and Garba XIII, a middle Acheulean site (*42*). Geomagnetic polarity timescale is from (*43*), $^{40}Ar/^{39}Ar$ dates are from (*5*), and K/Ar dates are from (*44*). (**C** and **D**) The stratigraphy of Garba IV and the paleomagnetic sequence of samples with normal (small black squares) and reverse polarity (small white squares) (*6*) (fig. S3).

LCTs constitute ~1% of the analyzed assemblage (large flakes, 41; LCTs, 23) but are relevant to understanding the technological shift from the underlying Oldowan assemblage (*8*). The features of the assemblage, including the production of large flakes and LCTs, identify it as Early Acheulean. Thus, Garba IV documents a rapid change between Oldowan and Acheulean lithic production over a period of 50 thousand to 100 thousand years (kyr) in this distinctive high-altitude paleoenvironment.

Paleoecology

The vegetation of the Garba IV site belonged to the dry evergreen Afromontane forest and grassland complex, which currently develops on the mountains of East Africa above 1800 m asl (*9*, *10*). The palynology indicates the presence of bushy woodlands alternating with patches of grassland in the area of Melka Kunture at around 1.8 Ma or possibly earlier. (*1*). A more open vegetation and a mountain grassland followed, which further expanded during the deposition of level D at Garba IV.

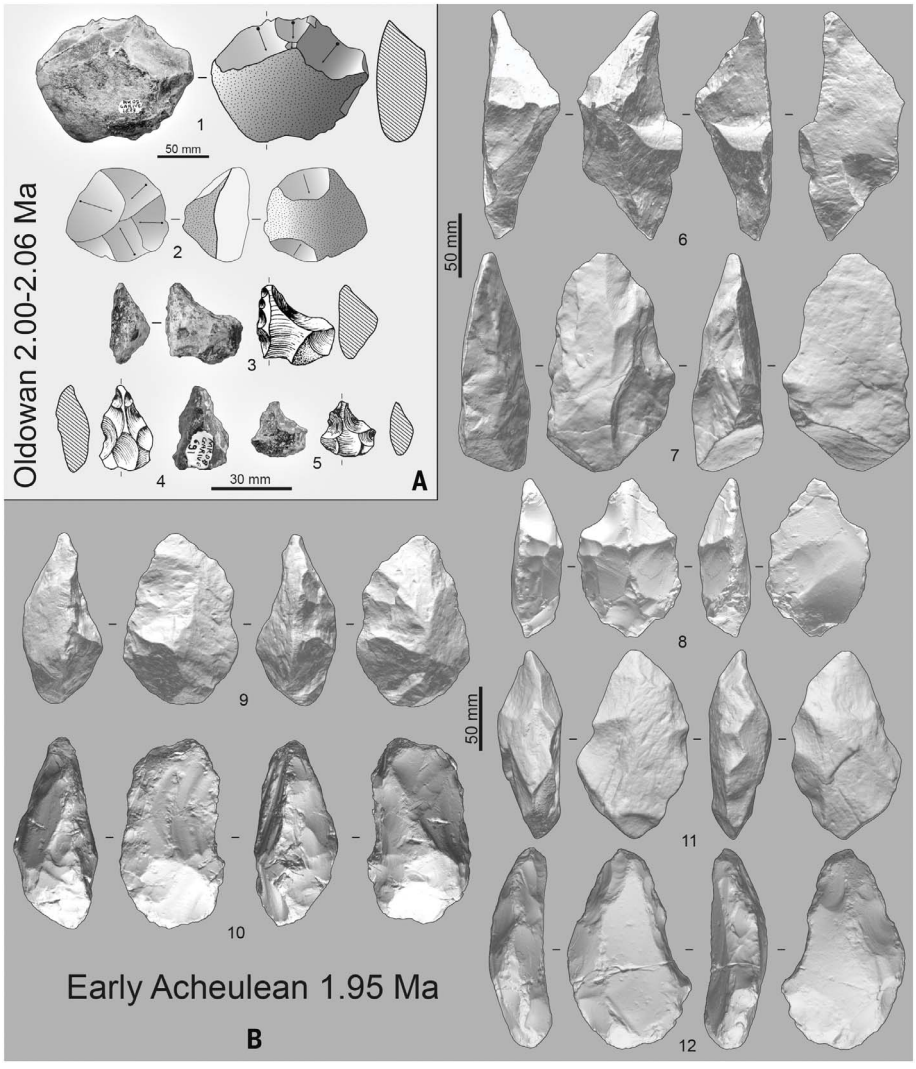


Fig. 2. Garba IV artifacts. (**A**) Levels E and F (2.00 Ma to 2.06 Ma), Oldowan. Images show obsidian cores (1 and 2) and obsidian pointed tools (3 to 5). (**B**) Level D (1.95 Ma), Early Acheulean. Images show large retouched flakes shaped on basalt or obsidian (6 to 8) and handaxes shaped on lava, basalt, or obsidian flakes (9 to 12).

Afromontane vegetation develops under a mountain climate with large diurnal temperature ranges, resulting in plant species that are distinct from those of the lowland savanna or woodland.

The mammalian faunal remains of levels E and F are highly fragmented, and species-level identifications are rare. The fossil sample from level D is larger (table S3), but the three assemblages cannot be ecologically distinguished. A rare occurrence in level E is the extinct civet *Pseudocivetta ingens*, also known from several eastern African sites between ~2.4 and ~1.5 Ma. *Theropithecus*, a genus quite rare at Melka Kunture, is represented by a single specimen from level D . In this layer, the ungulate fauna is dominated by large hippos and grazing bovids, of which almost 90% are alcelaphins and antilopins; this unambiguously points to an open environment and is consistent with the relatively high frequency of equids (*2*). Also consistent with an open environment are measurements of $^{13}C/^{12}C$ isotopic ratios of dental-enamel carbonates carried on hippopotamus (n = 11), bovid (n = 1), and equid (n = 1) teeth from level D. The results show that hippos, bovids, and equids consumed C_4 plants, suggesting the presence of C_4 high-elevation grasslands (*11*) (fig. S20).

The bovid material from levels E and F cannot be identified beyond generic rank, whereas level D alcelaphins (*Connochaetes* and *Damaliscus*) are close to those known from contemporaneous East African lower-elevation sites—up to 1500 m asl—but are distinct at least at the subspecific level. Thus, the fauna record testifies to some endemicity (*2*, *12*) of the Ethiopian highlands, a factor to be considered when dealing with hominin evolution and subsistence in this paleoenvironment.

Taxonomy of the Garba IVE mandible

Previous analyses of the external and internal morphology of the MK 81 GAR IVE 0043 (abridged name Garba IVE) mandible (Fig. 3) have identified affinities with various early *Homo* species such as *H. habilis*, *H. rudolfensis*, *H. ergaster*, and *H. erectus* (*13*). Although GAR IVE had been described as struck by a rare genetic disorder affecting its enamel (*amelogenesis imperfecta*) (*14*), this was recently refuted with synchrotron imaging, and features previously interpreted as pathological were instead shown to result from taphonomic processes (*15*). On the basis of medical computed tomography (CT) scans, Zanolli *et al*. (*3*) expanded on previous assessments to examine the sequence of dental development, tooth crown–tissue proportions, tooth size in the unerupted permanent dentition, and corpus cortical-bone thickness. Their conclusion was that the Garba IVE specimen shared affinities with *H. habilis/rudolfensis* and *H. erectus* sensu lato. We used synchrotron CT to produce high-resolution images of the enamel and dentine surfaces of the unerupted permanent dentition (*15*) because these have been shown to maximize the taxonomic information in hominin teeth (*16*–*18*). We conducted a three-dimensional landmark-based geometric morphometric analyses of the enamel-dentine junction of the third and fourth premolar and first molar. We were also able to compare the unerupted permanent dentition of Garba IVE against an expanded comparative sample that crucially includes specimens of *H. habilis* from both Tanzania and Kenya.

The enamel-dentine junction shape of the permanent third premolar exhibits morphological affinities with *H. erectus* that are based on similarities in crown height and the position and relative height of the metaconid (Fig. 4A). The enamel-dentine junction shape of the permanent first molar is intermediate between that of *Australopithecus* and later *Homo* and plots closest to KNM-ER 992. The shape reflects a number of features including a relatively reduced hypoconulid, which separates *H. erectus* from *H. habilis* (Fig. 4B). The enamel-dentine junction ridge of the permanent fourth premolar is less diagnostic [because the tooth is not crown complete (*15*), the cervix could not be used], and the GAR IVE specimen shows similarities to both *H. habilis* and *H. erectus* (fig. S21). Both premolars show strong similarities to KNM-WT 15000 at the outer enamel surface (fig. S22). The permanent canine is not crown complete [supplementary figure S1 in (*15*)], but the preserved enamel morphology is similar to that of *H. erectus* specimens such as KNM-ER 820 (fig. S22). The relative tooth-crown size of the premolars (i.e., having premolars that are similar in crown size) and first molar is also similar to that of *H. erectus* (Fig. 4C). It is currently difficult to make

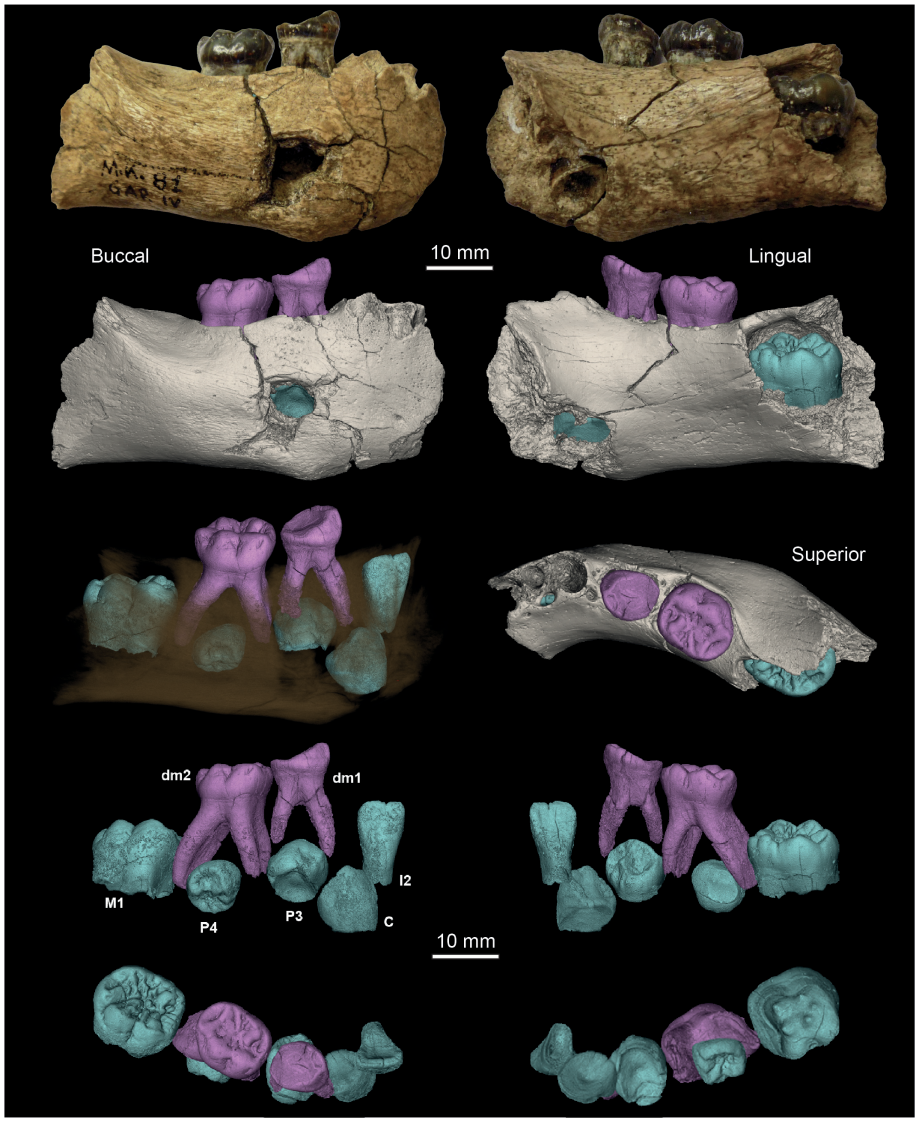


Fig. 3. The GAR IVE mandible. Views showing the preserved deciduous molars (dm1 and dm2) and the germs of the permanent teeth (I2, lateral incisor; C, canine; P3 and P4, third and fourth premolar; M1, first molar). The image was modified from (*15*).

comparisons with material definitively attributed to *H. rudolfensis* (e.g., KNM-ER 1482, 1801, and KNM-ER 60000) because of the limited number of specimens and the preservation and degree of attrition of many of the mandibular teeth, but we consider what is measurable of the dentition of both to be more similar to *H. habilis* than to *H. erectus*. Overall, this expanded analysis of the external and internal morphology of the Garba IVE specimen is consistent with an attribution to *H. erectus* rather than *H. habilis*.

Discussion

At ~2 Ma, the GAR IVE mandible is one of the earliest *H. erectus* fossils so far discovered and the only specimen whose taxonomic identification is based on teeth, which are known to have a strong taxonomic signal (*15*–*18*). Furthermore, it is the first to be directly associated in a sealed deposit with Oldowan stone tools in Africa. The same industry is also found in underlying level Garba IVF, pointing to a fully ingrained behavior that we attribute to *H. erectus* activity. The many small, pointed tools and the systematic use of obsidian make this assemblage conspicuously different from the 2.4- to 1.6-Ma Oldowan material found elsewhere, and notably at sites where remains of *Homo habilis* were also discovered (table S2). Six knapped stones classified as Oldowan have been reported at Drimolen (*19*), and although a ~2.04-Ma neurocranium (*20*) has been classified as *H.* aff. *erectus* on the basis of affinities with the Mojokerto cranium (and we note that there is no similarly aged *Paranthropus* cranium with which it can be compared), the site is dominated by *Paranthropus robustus* remains, and thus the hominin that produced the material cannot be determined with certainty (*19*, *21*). The innominate hip bone KNM-ER 3228 from Koobi Fora is not directly associated with stone tools (*22*), whereas at Olduvai PTK, the OH 86 proximal phalanx attributed to *Homo* cf. *H. erectus* sensu lato is associated with Oldowan stone tools (*23*) but is dated to 1.84 Ma. Hominin taxa such as *H. habilis*, *Paranthropus*, and *Australopithecus* lived at the low to middle elevation of eastern and southern Africa (up to 1500 m asl) and likely coexisted with *H. erectus*, such as at Drimolen, Olduvai, and Koobi Fora (*24*). On the basis of this study, however, only *H. erectus* is known to have inhabited the Ethiopian highland at or above 2000 m asl.

Around 2 Ma (i.e., before the Olduvai/Matuyama boundary at 1.925 Ma), there is a shift toward the production of Acheulean tools, such as handaxes and cleavers produced on large flakes. Previously, only statistical models had suggested such an early date (*25*), although published data based on fieldwork tentatively pointed to the initial development of this technology at ~1.8 Ma (*26*–*29*). The Garba IVD data suggest that the Acheulean actually emerged 200 kyr earlier and that there is a substantial chronological gap between the emergence of the Early Acheulean at Garba IV and in the Rift Valley. Compared with small-flake production in the Oldowan, this major step in technological and cultural development implies both the accurate search and management of larger blanks of stone and the strength to produce considerably larger flakes. Some large flakes were then partially retouched, and others were shaped into more complex tools. Outside Melka Kunture, fully knapped handaxes, with a true bilaterally and bifacially balanced shape, are documented 1.7 Ma at FLK West (*27*). At Melka Kunture, the Early Acheulean was also discovered at Gombore IB, later in age at ~1.66 Ma (*6*, *30*), where a very robust *Homo* humerus was found in association (*30*, *31*). As at Garba IVD, there were fully shaped handaxes, which implies that in the earliest phases of the Acheulean, the mental templates and skills of bifacial shaping were already developed.

The aridification of Africa, which started after 2.8 Ma (*32*, *33*), could have driven hominin groups beyond previously occupied environments. A highland environment, such as that at Garba IV, at an altitude of 2000 m asl and more, exposes individuals to less oxygen, higher exposure to ultraviolet rays, more rain, and cooler temperatures. Will *et al.* (*34*) have recently examined the *Homo* fossil record up to 1 Ma and show that Bergmann's rule applies, which predicts a larger body size in colder environments. The larger-bodied and larger-brained *H. erectus* was possibly better adapted to the highlands than were smaller-bodied hominins such as *Australopithecus* and *H. habilis*.

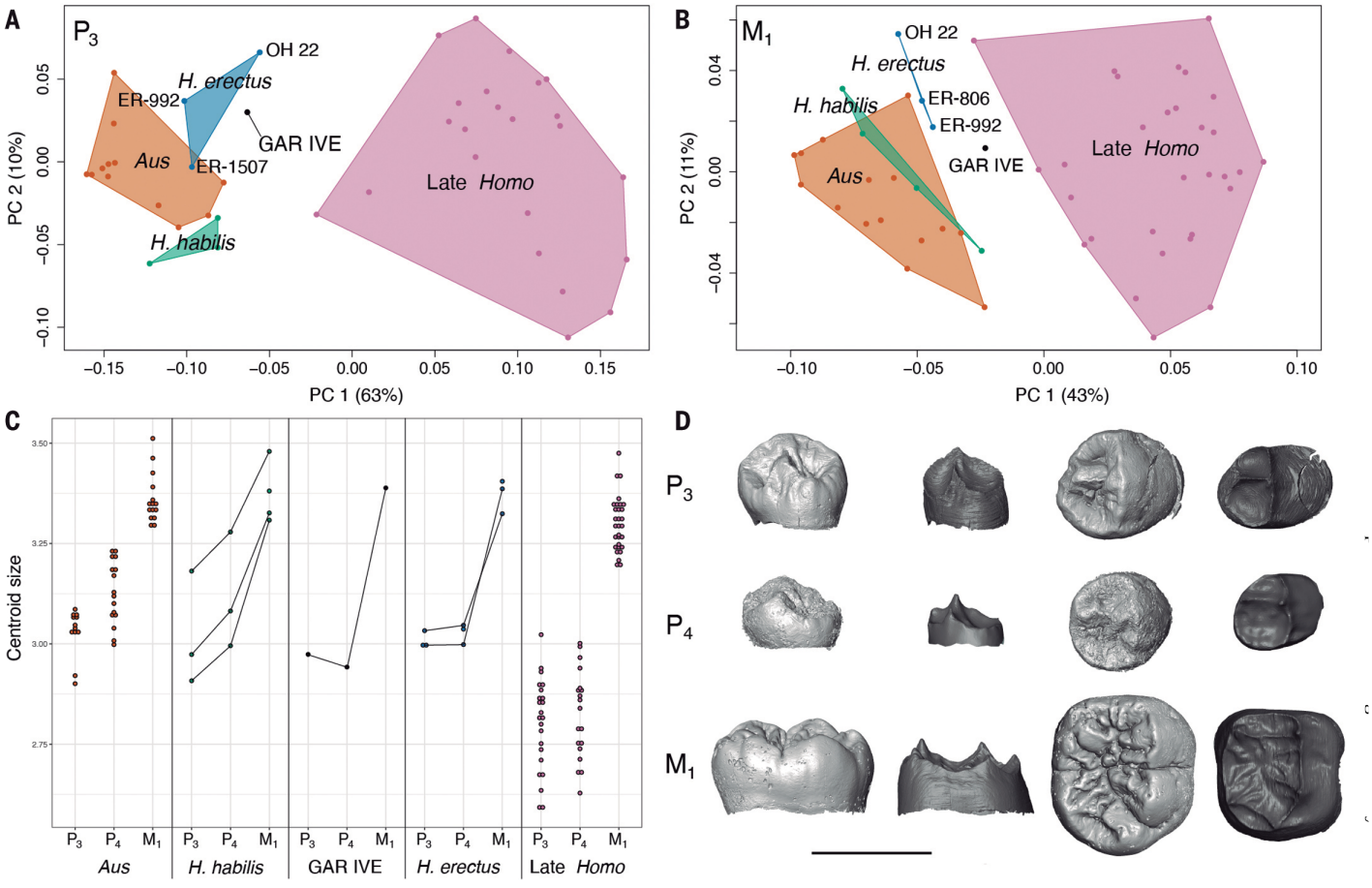


Fig. 4. Analysis of the internal tooth shape of the permanent postcanine dentition. (**A**) Plot of first and second principal components (PC 1 and PC 2) of mandibular third premolar enamel-dentine junction shape. (**B**) First and second principal components of mandibular first molar enamel-dentine junction shape. The variation explained by each principal component is shown in parentheses. In both principal components analysis (PCA) plots, GAR IVE is most closely associated with *H. erectus*. (**C**) Tooth-size variation (centroid size of enamel-dentine junction ridge landmarks) between the third premolar, fourth premolar, and first molar in GAR IVE and the comparative sample. (**D**) Surface models of the third premolar, fourth premolar, and first molar in (left) lingual and (right) occlusal view (enamel surface in light gray and enamel-dentine junction surface in dark gray).

In addition, the vegetation of Melka Kunture was different from that of the savanna, although some of its animal species testify of a degree of endemism. The hominins would have needed to acquire knowledge of these novel resources. Information on plant use is not preserved, but the nearby forests would have offered resources such as fruit, nuts, tubers, and roots. Butchery marks have been detected on hippopotamus and bovid bones from Garba IVD (*35*). The fossil and archaeological evidence from Garba IV indicates that *H. erectus* could cope with demanding local conditions, eventually changing their way of producing stone tools and configuring the pieces that define the emergence of the Acheulean technocomplex, as also recorded at Gombore IB.

The ~2-Ma core and flake assemblages made by *H. erectus* at Melka Kunture allow us to put into a new perspective the North African and Asiatic assemblages, which, later on, still include only small flakes without any handaxes or other large tools (*36*–*39*). These assemblages could be the outcome of an earlier expansion or the record of an expansion from areas where *H. erectus* and the Acheulean appeared later than in the Ethiopian highlands. In Sub-Saharan Africa, the early Acheulean of Garba IVD establishes that Oldowan and Acheulean industries coexisted over at least 300 kyr, which revives the debates about who were the makers of core and flake assemblages at lower elevations. At Olduvai (Bed II) and in the Middle Awash, the coexistence of different hominins producing different lithic industries has been suggested (*27*, *40*, *41*).

The hominin fossil record is dominated by the discoveries made at low to middle elevations in the Rift Valley and in the cave deposits of South Africa. The fossil and archaeological remains from Melka Kunture demonstrate that the highlands of East Africa, with their different paleoenvironment, are pivotal to understanding *H. erectus* behavior. Around 2 Ma, there is evidence at Garba IV of *H. erectus* retaining behavioral characteristics close to those of *H. habilis*. *H. erectus* was still producing core and flake assemblages but with different features than those of typical Oldowan technocomplexes. Then around 1.95 Ma, the Early Acheulean emerged with its archetypal bifacial tools. Between 2 and 1.9 Ma, the Melka Kunture site provides the earliest evidence of *H. erectus*, who quickly adjusted to a high-altitude environment, first producing Oldowan technology and then developing Acheulean technology.

REFERENCES AND NOTES

1. R. Bonnefille, R. T. Melis, M. Mussi, in *The Emergence of the Acheulean in East Africa and Beyond. Contributions in Honor of Jean Chavaillon*, R. Gallotti, M. Mussi, Eds. (Springer, 2018), pp. 93–114.
2. D. Geraads, V. Eisenmann, G. Petter, in *Studies on the Early Paleolithic Site of Melka Kunture, Ethiopia*, J. Chavaillon, M. Piperno, Eds. (Istituto Italiano di Preistoria e Protostoria, Roma, 2004), pp. 169–192.
3. C. Zanolli *et al.*, *Am. J. Phys. Anthropol.* **162**, 533–549 (2017).
4. E. Tamrat, N. Thouveny, M. Taieb, J. P. Brugal, *Quat. Int.* **343**, 5–16 (2014).

5. L. E. Morgan *et al*., *J. Hum. Evol.* **62**, 104–115 (2012).
6. S. Perini, G. Muttoni, E. Monesi, R. T. Melis, M. Mussi, *Quat. Sci. Rev.* **274**, 107259 (2021).
7. R. Gallotti, M. Mussi, *PLOS ONE* **10**, e0145101 (2015).
8. R. Gallotti, *J. Hum. Evol.* **65**, 594–620 (2013).
9. I. Friis, *Kew Bulletin Additional Series* XV (1992).
10. I. Friis, D. Sebsebe, *Biol. Skr.* **54**, 399–439 (2001).
11. S. H. Ascari, J. K. Njau, P. E. Sauer, P. D. Polly, Y. Peng, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **511**, 550–557 (2018).
12. D. Geraads, R. Gallotti, J. P. Raynal, R. Bonnefille, M. Mussi, in *African Paleoecology and Human Evolution,* S. Reynolds, R. Bobe, Eds. (Cambridge Univ. Press, 2022), pp. 256–268.
13. S. Condemi, in *Studies on the Early Paleolithic site of Melka Kunture, Ethiopia,* J. Chavaillon, M. Piperno, Eds. (Istituto Italiano di Preistoria e Protostoria, Roma, 2004), pp. 687–702.
14. U. Zilberman, P. Smith, M. Piperno, S. Condemi, *J. Hum. Evol.* **46**, 647–653 (2004).
15. A. Le Cabec *et al*., *Sci. Rep.* **11**, 23087 (2021).
16. T. W. Davies *et al*., *Sci. Rep.* **10**, 13196 (2020).
17. C. Zanolli *et al*., *Nat. Ecol. Evol.* **3**, 755–764 (2019).
18. M. M. Skinner, P. Gunz, B. A. Wood, J.-J. Hublin, *J. Hum. Evol.* **55**, 979–988 (2008).
19. R. C. Stammers, M. V. Caruana, A. I. R. Herries, *Quat. Int.* **495**, 87–101 (2018).
20. A. I. R. Herries *et al*., *Science* **368**, eaaw7293 (2020).
21. C. Zanolli *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2111212119 (2022).
22. R. E. F. Leakey, *Nature* **261**, 574–576 (1976).
23. M. Domínguez-Rodrigo *et al*., *Nat. Commun.* **6**, 7987 (2015).
24. R. Bobe, S. Carvalho, *J. Hum. Evol.* **126**, 91–105 (2019).
25. A. J. M. Key, D. L. Roberts, I. Jarić, *J. Hum. Evol.* **154**, 102976 (2021).
26. H. Duke, C. Feibel, S. Harmand, *J. Hum. Evol.* **159**, 103061 (2021).
27. F. Diez-Martín *et al*., *Sci. Rep.* **5**, 17839 (2015).
28. C. J. Lepre *et al*., *Nature* **477**, 82–85 (2011).
29. Y. Beyene *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1584–1591 (2013).
30. M. Mussi *et al*., *Quat. Int.* **657**, 3–25 (2023).
31. F. Di Vincenzo *et al*., *Quat. Sci. Rev.* **122**, 207–221 (2015).
32. P. B. deMenocal, *Earth Planet. Sci. Lett.* **220**, 3–24 (2004).
33. C. J. Lepre, R. L. Quinn, *Global Planet. Change* **208**, 103684 (2022).
34. M. Will, M. Krapp, J. T. Stock, A. Manica, *Nat. Commun.* **12**, 4116 (2021).
35. I. Fiore, A. Tagliacozzo, in *Studies on the Early Paleolithic site of Melka Kunture, Ethiopia*, J. Chavaillon, M. Piperno, Eds. (Istituto Italiano di Preistoria e Pratostoria, Florencia, 2004), pp. 639–682.
36. G. Scardia, W. A. Neves, I. Tattersall, L. Blumrich, *Evol. Anthropol.* **30**, 122–127 (2021).
37. M. Sahnouni *et al*., *Science* **362**, 1297–1301 (2018).
38. Z. Zhu *et al*., *Nature* **559**, 608–612 (2018).
39. R. Ferring *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10432–10436 (2011).
40. K. Schick, N. Toth, *Anthropologie* **121**, 451–491 (2017).
41. M. D. Leakey, *Olduvai Gorge. Vol. 3, Excavations in Beds I and II, 1960-1963*. (Cambridge Univ. Press, 1971).
42. R. Gallotti, J.-P. Raynal, D. Geraads, M. Mussi, *Quat. Int.* **343**, 17–27 (2014).
43. J. E. T. Channell, B. S. Singer, B. R. Jicha, *Quat. Sci. Rev.* **228**, 106114 (2020).
44. J.-J. Schmitt, J.-M. Wempler, J. Chavaillon, M. C. Andrews, "Initial K/Ar and Paleomagnetic Results of the Melka-Kunturé early-man sites, Ethiopia" in *Proceedings of the 8th Panafrican Congress of Prehistory and Quaternary Studies, Nairobi, 5 to 10 September 1977*, R. E. Leakey, B. A. Ogot, Eds. (Louis Leakey Memorial Institute for African Prehistory, 1980).

ACKNOWLEDGMENTS

We are grateful to the Authority for Research and Conservation of the Cultural Heritage (currently the Ethiopian Heritage Authority) and to the Ministry of Culture and Tourism of the Federal Republic of Ethiopia for the loan and temporary export permit of MK 81 GAR IVE 0043 (27 November 2015 to 26 January 2016) awarded to M. Mussi. For access to comparative fossil specimens (and associated CT scans), we thank the following individuals and institutions: B. Zipfel and S. Jira (Evolutionary Studies Institute, University of the Witwatersrand); E. Mbua and J. Kibii (National Museums of Kenya); A. Mabulla (National Museum of Tanzania); the Tanzania Commission for Science and Technology; M. Endalemaw and Y. Assefa (Ethiopian Authority for Research and Conservation of Cultural Heritage); W. Kimbel (Arizona State University); Z. Alemseged (University of Chicago); F. Spoor (Natural History Museum, London); J.-F. Tournepiche (Musée d'Angoulême); J.-J. Cleyet-Merle (Musée National de Préhistoire, Les Eyzies-de-Tayac); J. Radovčić (Croatian Natural History Museum); F. Groening (Senckenberg, Forschungsstation für Quartärpaläontologie); V. Merlin-Langlade (Musée d'Art et d'Archéologie du Périgord); and M. Toussaint (ASBL Archéologie Andennaise and Royal Belgian Institute of Natural Sciences). Isotopic analyses have been funded by the Biogeology Lab (University of Tübingen), and we thank P. Tung for technical support. For technical assistance, we thank F. Spoor, D. Plotzki, and H. Temming. We are grateful to P. Tafforeau and colleagues at the ESRF for accessing beamtime on the ID19 beamline and for technical assistance. This research benefited from the scientific framework of the University of Bordeaux's IdEx Investments for the Future program/GPR Human Past. **Funding:** This research has been funded by Università di Roma Sapienza grant C26S133Z4B (M.M.); Università di Roma Sapienza grant SA11715C7C936C01 (M.M.); Ministero degli Affari Esteri e della Cooperazione Internazionale grant ARC-000799 (M.M.); Ministero degli Affari Esteri e della Cooperazione Internazionale grant ARC-001666 (M.M.); European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement 819960 (M.M.S.); Fundación Palarq (Spain) grants 2019 (J.P. and M.M.), 2020 (J.P., M.M., E.M.-Q., and S.R.-J.), and 2021 (J.P., M.M., E.M.-Q., and S.R.-J.); Leakey Foundation grant 42137 (E.M.-Q.). Deutscher Akademischer Austauschdienst (DAAD) grant 57552336 (G.B.); Università di Roma Sapienza grant 0044296 (G.B.); CNRS (A.L.C.); and postdoc Xunta de Galicia grant ED481D-2022/023 (E.M.-Q.). **Author contributions:** Conceptualization: M.M. Methodology: L.D.B. reopened the excavated area and reidentified the archaeological levels for sampling. G.B. collected the faunal teeth-enamel samples (sampling and pretreatment). Investigation: R.T.M., stratigraphy and geology; M.M.S., T.W.D., A.G., A.L.C., and J.-J.H., analysis of the dental samples; D.G., paleontology; H.B. and G.B., stable isotopes; E.M.-Q., J.P., and S.R.-J., archaeological analysis; R.B., paleobotany; and A.L.C., analysis of the synchrotron micro-CT scans. Funding acquisition: M.M., M.M.S., J.P., E.M.-Q., S.R.-J., G.B., and A.L.C. Project administration: M.M., J.P., and E.M.-Q. Supervision: M.M. Writing – original draft: M.M., with contributions from all authors. Writing – review and editing: M.M. and E.M.-Q., with contributions from all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data supporting the findings of this study are available within the paper and its supplementary materials. Landmark data and R code used to conduct the geometric morphometric analyses used in this study are available on the Publications page of Human Fossil Record.org (https://human-fossil-record.org/index.php?/category/595). The hominin mandible MK 81 GAR IVE 0043, as well as the paleontological and archaeological collections from the site, are kept at the National Museum in Addis Ababa and are available for scientific research. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add9115
Materials and Methods
Figs. S1 to S22
Tables S1 to S4
References (*45–100*)
MDAR Reproducibility Checklist

Submitted 15 July 2022; resubmitted 16 December 2022
Accepted 29 September 2023
10.1126/science.add9115

PLANT SCIENCE

Plant cell wall patterning and expansion mediated by protein-peptide-polysaccharide interaction

Steven Moussu[1]†‡, Hyun Kyung Lee[1]†, Kalina T. Haas[2], Caroline Broyart[1], Ursina Rathgeb[3], Damien De Bellis[3,4], Thomas Levasseur[5], Sébastjen Schoenaers[2,6], Gorka S. Fernandez[7], Ueli Grossniklaus[7], Estelle Bonnin[5], Eric Hosy[8], Kris Vissenberg[6,9], Niko Geldner[3], Bernard Cathala[5], Herman Höfte[2]*, Julia Santiago[1]*

Assembly of cell wall polysaccharides into specific patterns is required for plant growth. A complex of RAPID ALKALINIZATION FACTOR 4 (RALF4) and its cell wall–anchored LEUCINE-RICH REPEAT EXTENSIN 8 (LRX8)–interacting protein is crucial for cell wall integrity during pollen tube growth, but its molecular connection with the cell wall is unknown. Here, we show that LRX8-RALF4 complexes adopt a heterotetrametric configuration in vivo, displaying a dendritic distribution. The LRX8-RALF4 complex specifically interacts with demethylesterified pectins in a charge-dependent manner through RALF4's polycationic surface. The LRX8-RALF4-pectin interaction exerts a condensing effect, patterning the cell wall's polymers into a reticulated network essential for wall integrity and expansion. Our work uncovers a dual structural and signaling role for RALF4 in pollen tube growth and in the assembly of complex extracellular polymers.

A central question in plant biology is how cell wall polymers assemble into specific mesoscale patterns, conferring to the walls their physicochemical properties that allow growth and development. Plant cell walls consist of cellulose microfibrils and matrix polymers, which include the polyanionic pectins. Among the pectins, homogalacturonan is the most abundant polymer in expanding cell walls (*1*). During cell expansion, regulation of homogalacturonan charge through demethylesterification plays a key role in controlling cell wall mechanics (*2*–*5*). Cell expansion needs to be tightly controlled by rapid feedback signaling loops that coordinate cell wall deposition and remodeling to avoid loss of cell wall integrity. Here, we show that RAPID ALKALINIZATION FACTOR 4 (RALF4) peptides are required for cell wall patterning through their interaction with demethylesterified homogalacturonan. RALFs were previously identified as signaling peptides that coordinate the functional integrity of cell walls in a plethora of processes underlying plant development and innate immunity (*6*–*14*). These peptides instruct, in a mutually exclusive way and with different affinities, two structurally distinct protein families, the cell wall–anchored LEUCINE-RICH REPEAT EXTENSIN (LRX) proteins (*9*, *10*, *15*) and a membrane-integral signaling complex composed of LORELEI-like GLYCOLPHOSPHATIDYLINOSITOL (GPI)-ANCHORED PROTEINS (LLGs) and *Catharanthus roseus* RLK1-like receptor kinases (*Cr*RLK1Ls) (*7*, *12*, *14*, *16*). In this study, we used the pollen tube of *Arabidopsis thaliana*, which has a rapidly growing tip, as a model system to investigate the molecular links among RALFs, LRX proteins, and the cell wall matrix during cell expansion.

[1]The Plant Signaling Mechanisms Laboratory, Department of Plant Molecular Biology, University of Lausanne, 1015 Lausanne, Switzerland. [2]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000 Versailles, France. [3]Department of Plant Molecular Biology, University of Lausanne, 1015 Lausanne, Switzerland. [4]Electron Microscopy Facility, University of Lausanne, 1015 Lausanne, Switzerland. [5]INRAE, UR1268 BIA, F-44300 Nantes, France. [6]Integrated Molecular Plant Physiology Research (IMPRES), Department of Biology, University of Antwerp, 2020 Antwerp, Belgium. [7]Department of Plant and Microbial Biology & Zurich-Basel Plant Science Center, University of Zurich, 8008 Zurich, Switzerland. [8]IINS, CNRS UMR5297, University of Bordeaux, 33000 Bordeaux, France. [9]Plant Biochemistry & Biotechnology Lab, Department of Agriculture, Hellenic Mediterranean University, Stavromenos PC 71410, Heraklion, Crete, Greece.
*Corresponding author. Email: hermanus.hofte@inrae.fr (H.H.); julia.santiago@unil.ch (J.S.)
†These authors contributed equally to this work.
‡Present address: Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, Toulouse INP, 31320 Auzeville Tolosane, France.

LRX-bound RALF4 interacts with charged homogalacturonan

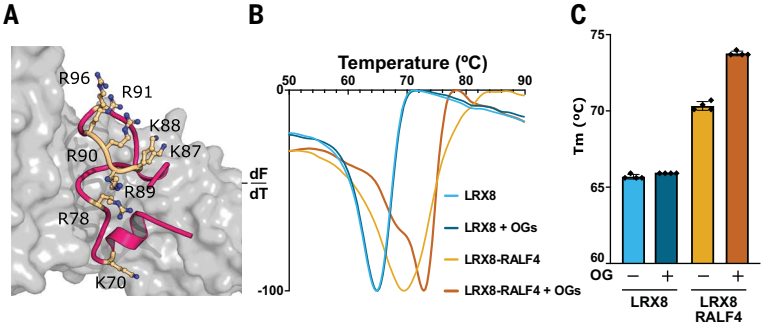
Our previous structural analysis of the LRX8-RALF4 signaling complex in the pollen tube revealed that RALF4 peptides, when bound to LRX8, expose an alkaline surface patch rich in Lys and Arg residues (*10*) (Fig. 1A). We hypothesized that this surface patch could serve as a platform to interact with negatively charged cell wall polysaccharides such as pectins. The pollen tube grows inside the pistil toward the ovule and must navigate through various tissues, counterbalancing different physical forces (*17*, *18*). Along this journey, the cell wall at the pollen tube apex constantly needs to adjust its mechanical properties to allow rapid growth while avoiding its rupture (*2*, *3*, *19*). Pectin charge plays a key role in this process (*3*–*5*).

We initially compared the melting profile of recombinantly expressed LRX8 and LRX8-RALF4 complexes (*10*) in the presence and absence of fully demethylesterified homogalacturonan fragments (6- to 15-mer oligogalacturonans hereafter referred as OGs) in thermal shift assays (TSAs) (*20*) (Fig. 1, B and C, and fig. S1). Consistent with our isothermal titration calorimetry (ITC) experiments, these assays revealed an increase in melting temperature (T_m) of LRX8 upon binding to RALF4 (Fig. 1, B and C) (*10*). The T_m further increased in the presence of OGs, whereas no such changes were observed for LRX8 alone (Fig. 1, B and C, and fig. S1). In addition, the LRX8-RALF4 complex did not bind shorter homogalacturonan fragments (fig. S1), a structurally divergent acidic polysaccharide (polymannuronic acid), a negatively charged polymer (DNA), or an unrelated carbohydrate (sucrose) (Fig. 1D and fig. S1). We next quantified these interactions by titrating OGs into an LRX8-RALF4 solution (*10*) using ITC (*21*). LRX8-RALF4 bound OGs at ~2 μM affinity and a binding stoichiometry (N) of 2, suggesting that one OG molecule binds to each of the two RALF4 molecules in the heterotetrametric (2 + 2) LRX8-RALF4 complex (*10*) (Fig. 1E and fig. S2). Consistently, a monomeric version of the complex, monomeric LRX8-RALF4 (*10*), bound one OG molecule (N = 1) with comparable affinity (~2.9 μM), and its T_m increased in the presence of OGs (Fig. 1E and fig. S2). LRX2, a protein that is not expressed in pollen tubes, in complex with RALF4, bound OGs with a similar dissociation constant (1.6 μM) and stoichiometry (Fig. 1E and fig. S2). No binding was detected when titrating OGs into a solution containing either LRX8 alone or free RALF4 or when titrating DNA into the LRX8-RALF4 complex (Fig. 1E and fig. S2). These experiments demonstrate that the LRX-RALF4 complex can bind homogalacturonan oligomers with high affinity and specificity.

During pollen tube growth, demethylesterification at the apex changes the overall electrostatic charge of homogalacturonan, which plays a role in controlling cell wall expansion (*3*, *5*, *22*). The degree of methylation (DM) of homogalacturonan was critical for the interaction with LRX8-RALF4, as shown by its increased thermostability in the presence of non- or low-methylesterified (DM5 = 5% methylesterified) OGs (increase in T_m of 3.9° and 2.5°, respectively), but not of highly methylesterified ones (DM71) (Fig. 1F and fig. S2). Similar results were obtained when using longer pectin chains: The shift in T_m for DM20 pectin occurred at lower concentrations compared with DM43 pectins (Fig. 1G and fig. S2). These results show that LRX8-RALF4 specifically interacts with demethylesterified pectins.

To understand the interaction between LRX8-RALF4 and demethylesterified pectins in molecular terms, we mapped and probed the basic surface of the LRX8-bound RALF4 peptide (*10*) (Fig. 1A). We mutated residues

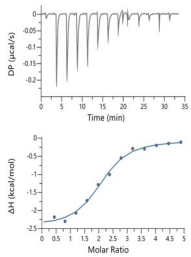
Fig. 1. LRX-bound RALF4 specifically interacts with demethylesterified homogalacturonan in a charge-dependent manner. (**A**) Structural representation of the polycationic interface of RALF4 when bound to LRX8 [Protein Data Bank identifier: 6QWN (*10*)]. LRX8 is depicted in surface representation (gray) and RALF4 as a pink ribbon diagram. Positively charged amino acids in the exposed RALF4 interface are depicted as yellow sticks. (**B**) LRX8-bound RALF4 binds demethylesterified homogalacturonan. Representative thermal denaturation profiles of LRX8 and LRX8-RALF4 in the absence and presence of OGs. (**C**) Graphical representation of T_m of different TSAs shown in (B). Data are shown as means ± SD of four independent experiments. (**D**) Summary table of T_m and ΔT_m of LRX8 and LRX8-RALF4 in the presence of OGs and control molecules (n = 4). (**E**) ITC thermogram of LRX8-RALF4 versus OGs (left) and ITC summary table (right) of LRX8 alone, RALF4, and LRXs-RALF4 versus OGs and DNA oligomers. K_d (dissociation constant) indicates the binding affinity between the two molecules considered (in micromolar). N is the reaction stoichiometry (N = 1 for a 1:1 interaction). The values indicated in the table are shown as means ± SD of at least two independent experiments. n.d., no detectable binding. (**F**) Summary table (top) and representation of T_m (bottom) of LRX8-RALF4 versus OGs and OGs with different degrees of methylesterification (e.g., DM71 = 71% methylesterification). Data are shown as means ± SD; n = 4. (**G**) T_m bar representation of LRX8-RALF4 in the presence of increasing concentrations of long chains of homogalacturonan with DM43 and DM20. Data are shown as means ± SD; n = 4. (**H**) Summary table of the T_ms shown in (G). (**I** and **J**) Mutation of RALF4 polycationic surface disrupts the interaction with homogalacturonan. (I) Summary table of T_ms of RALF4$_{R89E.R90E.R91E}$ and RALF4$_{neutral\ surface}$ variants in the presence of OGs. (J) ITC of RALF4$_{neutral\ surface}$ versus OGs (left). (Right) ITC summary table of RALF4$_{R89E.R90E.R91E}$ and RALF4$_{neutral\ surface}$ mutants versus OGs. The values indicated in the table are shown as the means ± SD of at least two independent experiments.



D

		Mock	OGs	Mannuronic Acid	DNA	Sucrose
LRX8	Tm	65.7	65.9	66.2	66.0	66.0
	ΔTm		0.2	0.5	0.3	0.3
LRX8-RALF4	Tm	70.3	73.8	71.2	69.7	69.8
	ΔTm		3.4	0.9	-0.7	-0.5

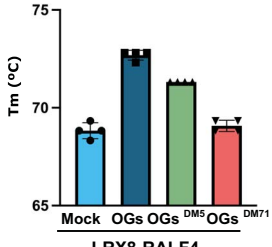
E



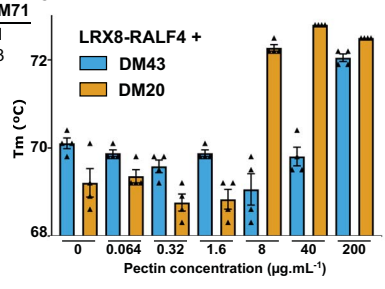
Ligand (syringe)	Protein (cell)	Kd (µM)	ΔH (kcal/mol)	N
OGs	LRX8-RALF4	2 ± 0.5	-2.5 ± 0.1	2
OGs	mLRX8-RALF4	2.9 ± 0.4	-1.8 ± 0.6	1
OGs	LRX8	n.d		
OGs	RALF4	n.d		
OGs	LRX2-RALF4	1.6 ± 0.2	-1.4 ± 0.11	2
DNA	LRX8-RALF4	n.d		

F

		Mock	OG	OG-DM5	OG-DM71
LRX8-RALF4	Tm	68.8	72.7	71.3	69.1
	ΔTm		3.9	2.5	0.3



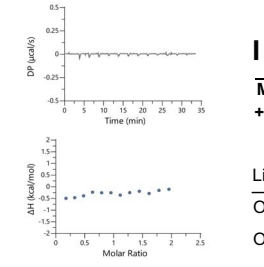
G



H

Ligand		Pectin concentration (µg/mL) 0	0.064	0.32	1.6	8	40	200
DM43	Tm	69.2	69.9	69.6	69.9	69.1	69.8	72.0
	ΔTm		0.7	0.4	0.7	-0.1	0.6	2.8
DM20	Tm	70.1	69.4	68.8	68.8	72.3	72.8	72.5
	ΔTm		-0.7	-1.3	-1.3	2.2	2.7	2.4

J



I

		LRX8-RALF4$_{R89E.R90E.R91E}$	LRX8-RALF4$_{neutral\ surface}$
Mock	Tm	73.7	73.4
+ OGs	Tm	74.6	72.1
	ΔTm	**0.9**	**-1.3**

Ligand (syringe)	Protein (cell)	Kd (µM)	ΔH (kcal/mol)	N
OGs	LRX8-RALF4$_{R89E.R90E.R91E}$	8.5 ± 0.4	-5.6 ± 1.6	2
OGs	LRX8-RALF4$_{neutral\ surface}$	n.d.		

in RALF4's basic patch to either Ala or Glu, and recombinantly expressed two mutant variants of the LRX8-RALF4 complex, one targeting the protruding loop of the folded signaling peptide (RALF4$_{R89E.R90E.R91E}$) and the other neutralizing the basic surface altogether (RALF4$_{K70A.R78A.K87A.K88A.R89A.R90A.R91A.R96A}$; hereafter referred as RALF4$_{neutral\ surface}$) (Fig. 1A and fig. S3). Both RALF4 mutants still bound LRX8 and LLG3 with wild-type–like affinity (fig. S3) (*10*). However, TSAs failed to detect an OG-induced increase in T_m for both LRX8-bound RALF4 mutants (Fig. 1, B to D and I, and fig. S4). Consistent with these findings, LRX8-RALF4$_{neutral\ surface}$ and LRX8-RALF4$_{R89E.R90E.R91E}$ showed either no or about a fourfold reduced OG-binding relative to the wild-type complex in ITC assays, respectively (Fig. 1, E and J, and fig. S4). Together, these results confirm the electrostatic nature of the interaction between LRX8-RALF4 and pectin.

LRX8, RALF4, and pectin form a reticulated pattern in vivo

Next, we investigated whether LRX8, RALF4, and pectin also interact in vivo. To this end, we performed immunolabeling and high-resolution

microscopy on pollen tubes expressing functionally tagged RALF4-HA and LRX8-myc under their native promoters (fig. S5) (*10*). Consistent with our structural and biochemical data (*10*), both proteins colocalized in pollen tube cell walls, suggesting that the complex is also present in vivo. However, the complex was present all along the pollen tube cell wall (Fig. 2A and fig. S6), not just at the tip, where the peptide and LRX8 are secreted. This indicates that RALF4, like LRX8, become an integral part of the pollen tube cell wall, which is an unexpected finding for a signaling peptide. This was confirmed by dSTORM super-resolution microscopy (*23*, *24*), which revealed a comparable distribution for the RALF4-HA peptide and LRX8-myc in punctate domains forming a reticulated network (Fig. 2, B and D). Moreover, cluster-size analysis of LRX8-myc and RALF4-HA was consistent with the presence of either one or two copies of the tags in each puncta, indicating the presence of heterotetrametric complexes in vivo (Fig. 2, C and E) (*10*). Concerning the distribution of demethylesterified homogalacturonan, Airyscan and dSTORM microscopy showed that the LM19 immunolabel for demethylesterified pectin

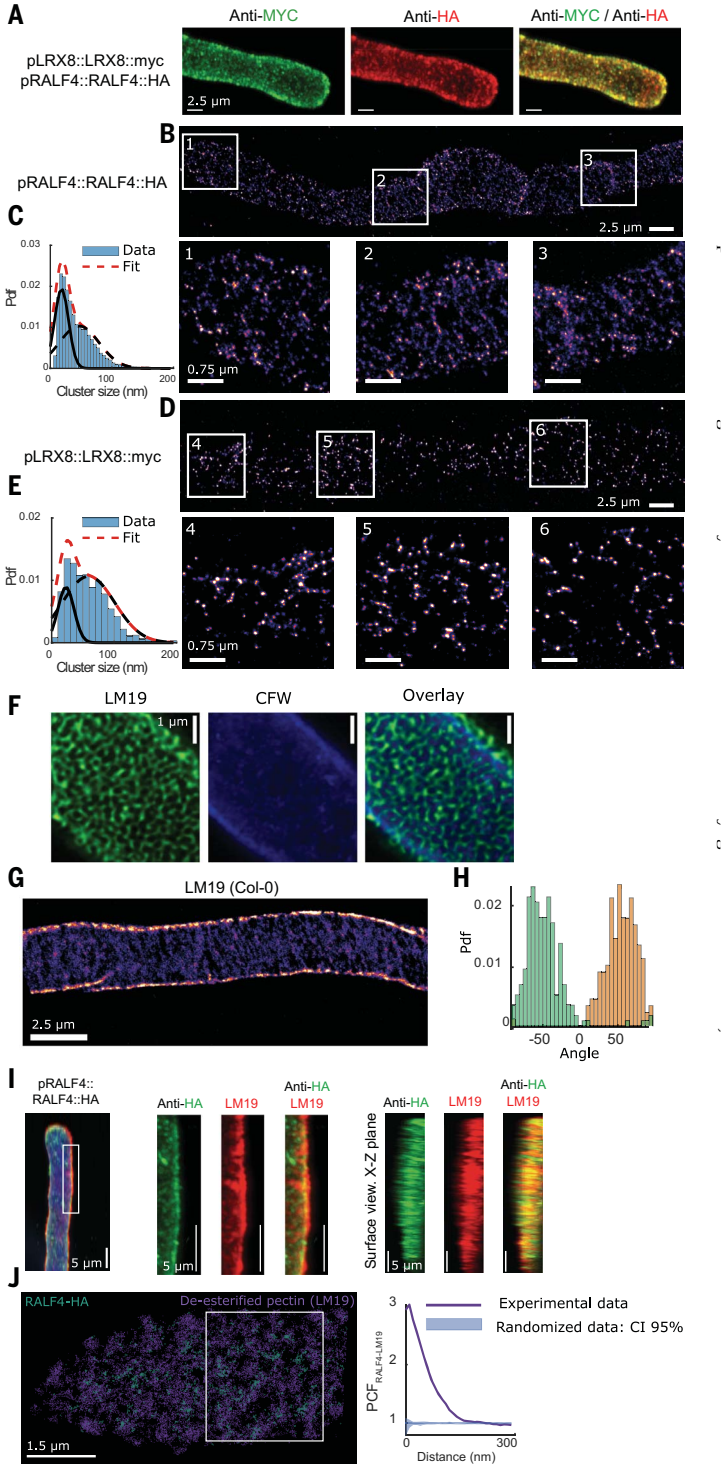


Fig. 2. LRX8, RALF4, and pectin associate in vivo and form a reticulated pattern. (**A**) Representative immunolabeled pollen tube images of a transgenic line expressing LRX8-myc and RALF4-HA under their native promoter in the Col-0 accession. (**B**) Representative pollen tube dSTORM image of immunolabeled *pRALF4::RALF4::HA* in Col-0. Bottom: Magnification of the corresponding squared regions marked in the pollen tube shank shown above. (**C**) Histogram of RALF4-HA probability density function (Pdf) versus cluster size (blue). Fit of the Gaussian mixture model shows two cluster populations with a mean size of 18 nm (monomer 35%) and 47 nm (dimers 65%). A total of 1137 clusters were analyzed from seven pollen tubes from four plants. (**D**) Top: Representative pollen tube dSTORM image of immunolabeled *pLRX8::LRX8::myc* in Col-0 background. Bottom: Magnification of the corresponding squared regions marked in the pollen tube shank shown above. (**E**) Histogram of LRX8-myc Pdf versus cluster size (blue). Two cluster populations of Gaussian mixture shows mean sizes of 23 nm (monomers 46%) and 59 nm (dimers 54%). A total of 937 clusters were analyzed from seven pollen tubes from four plants. (**F**) Representative orthogonal projection of Airyscan *z* stacks of demethylesterified homogalacturonan labeled with LM19 in wild-type pollen tubes. Calcofluor white (CFW) signal, mainly staining callose, did not show a patterned distribution. (**G**) Representative dSTORM image of a wild-type pollen tube labeled with LM19. (**H**) Histogram of Pdf versus angle, showing two distinct populations of pectin filaments (n = 4). (**I**) Confocal images of sequential immunolabeling targeting *pRALF4::RALF4::HA* in Col-0, and LM19 in pollen tube shanks. Maximal intensity *Z*-projected image (middle) and pollen tube surface view of the *X*–*Z* plane (right). (**J**) Left: Representative pollen tube dSTORM image with immunolabeled *pRALF4:RALF4-HA* and LM19. Right: Bivariate point correlation function (PCF) analysis calculated between RALF4-HA (green) and LM19 (violet) compared with the PCF calculated on randomized datasets. The PCF of experimental data lies above the 95% randomization confidence interval (CI), indicating spatial association between the protein and the polysaccharide.

(*25*) was not distributed randomly, as would be expected for an amorphous pectin gel, but rather formed an interconnected pattern of fibers (Fig. 2, F and G, and figs. S7 and S8) (*26*, *27*). This observed architecture is reminiscent of previously described ring like pectin bands in pollen tubes (*28*, *29*). The pattern presented a bimodal angle distribution, displaying as LRX8-RALF4, a web-like network (Fig. 2, F to H, and figs. S7 and S8). The interaction of LRX8-RALF4 with demethylesterified homogalacturonan in vivo was corroborated by the colocalization of RALF4-HA with LM19 antibodies in a periodic banding pattern at the pollen tube surface (Fig. 2I). This was confirmed at higher resolution by dSTORM microscopy, in which a two-point correlation function analysis showed that domains containing LM19-labeled pectin and RALF4-HA were not distributed independently of each other, but rather were associated in space, exhibiting a reticulated pattern (Fig. 2J and fig. S9).

Cell wall patterning and integrity through LRX8-RALF4-pectin compaction

The reticulated pattern of the LM19-labeled pectin and RALF4-HA domains could indicate the association of LRX8-RALF4 with preexisting fibrous structures or a more active role of the complex in the patterning of the cell wall polysaccharides. To investigate the latter possibility, we first studied in vitro how exposure to LRX8-RALF4 affects the rheology of a pectin layer using quartz-crystal microbalance with dissipation monitoring (QCM-D) (*21*, *30*). Briefly, to prepare the pectin layer, a pectin solution (DM43) was poured over an oscillating quartz crystal coated with a polycationic polymer while monitoring the frequency (ΔF, negative for a mass increase) and dissipation rate of the oscillation upon interruption of the current (ΔD, positive for increase in viscoelasticity of the layer). The drop in ΔF and increase in ΔD indicated the formation of an elastic pectin layer (Fig. 3A and fig. S10). The addition of free RALF4 induced an increase in ΔF rather than a decrease, which would be expected for the binding of the peptide, and a strong decrease in ΔD. This is consistent with an electrostatic interaction between the cationic RALF4 and anionic pectin, causing the hydrated pectin layer to dewater and stiffen. The increase in ΔF indicates that the mass loss due to dewatering exceeds the mass gain due to the binding of the small 5.6-kDa peptide. The RALF4-pectin interaction was reversible, as shown by the slow reversal of ΔF and ΔD during a subsequent washing step (Fig. 3B and fig. S10). Addition of LRX8-RALF4 to the pectin layer resulted in an amplified decrease in ΔF and ΔD (Fig. 3B). Here, the mass increase of the much larger complex (110 kDa) exceeded the mass of the expulsed water, explaining the decrease in ΔF. The complex, in contrast to the free RALF4 peptide, had a lasting condensing effect on the pectin layer, because ΔF and ΔD did not reverse during the washing step (Fig. 3B). Consistent with our previous binding assays, LRX8 alone did not induce any changes either in frequency (binding) or pectin rigidification, confirming that RALF4 targets the complex to pectins. The

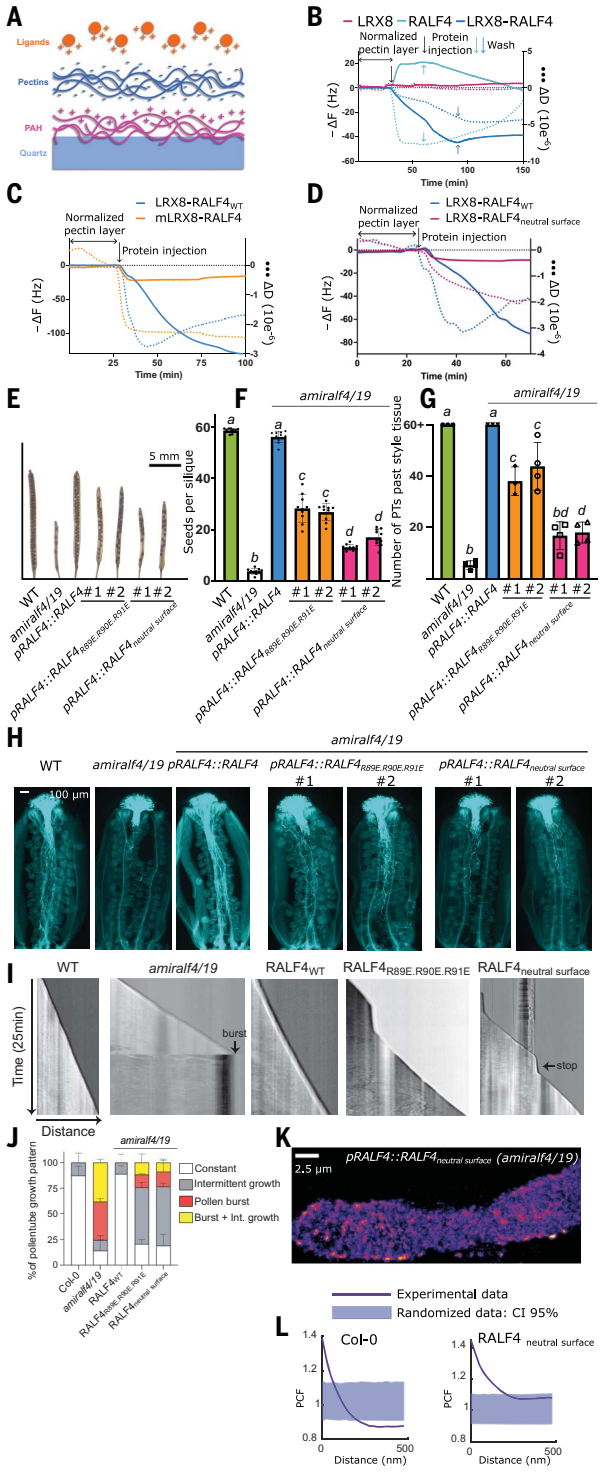


Fig. 3. LRX8-RALF4–induced pectin compaction promotes cell wall patterning regulating pollen tube growth. (**A**) Schematic overview of the QCM-D binding experiments (fig. S10). (**B** to **D**) Representative QCM-D profiles corresponding to the frequency (ΔF, solid lines), and dissipation (ΔD, dotted lines). Protein was injected over a DM43 pectin layer during the period indicated by the arrowheads. (B) QCM-D representative profiles of injecting RALF4 (light blue) (n = 7), LRX8-RALF4 (dark blue) (n = 16), and LRX8 (pink) (n = 3). (C) Representative profiles of dimeric LRX8-RALF4 (blue) (n = 16) and monomeric LRX8-RALF4 (orange) (n = 3). (D) Representative profiles of LRX8-RALF4 (blue) (n = 16), and LRX8-RALF4 $_{\text{neutral surface}}$ (pink) (n = 4). (**E** and **F**) Seeds per silique of Col-0 (wild-type, WT), *amiralf4/19*, *pRALF4::RALF4*$_{WT}$, and RALF4 mutants. Error bars are ± SD; n = 3. (**G**) Number of survived pollen tubes past style tissue. Error bars are ± SD; n = ~3 to 4. $P < 0.0001$, one-way ANOVA with Tukey's multiple-comparisons HSD test. (**H**) Representative images of pollinated *Arabidopsis* pistils stained with aniline blue. Pollen from Col-0, *amiralf4/19*, *pRALF4::RALF4*$_{WT}$, and RALF4 mutants was applied to Col-0 stigmas. (**I**) Representative images of growth kymographs. Arrows indicate either pollen tube bursting or stopping. (**J**) Quantification of growth patterns from kymographs shown in (I). Error bars are ± SD, 15 pollen tubes each; n = 2. (**K**) Representative dSTORM image of LM19 on pollen tubes from *amiralf4/19* complemented with *pRALF4::RALF4*$_{\text{neutral surface}}$. Data are represented as a two-dimensional pixelated image (pixel size, 15 nm). (**L**) Univariate PCF analysis calculated on LM19 dSTORM pointillist data on pollen tubes from Col-0 (left, representative image in Fig. 2G) and *amiralf4/19* complemented with *pRALF4::RALF4*$_{\text{neutral surface}}$ (right). PCF analysis of the distribution of the LM19 epitope shows that the signal crosses the CI in wild-type PTs at shorter distances relative to *pRALF4::RALF4*$_{\text{neutral surface}}$. This indicates that the LM19 epitope in wild-type pollen tubes forms densely packed, spatially segregated fibers compared with the *RALF4*$_{\text{neutral surface}}$ variant.

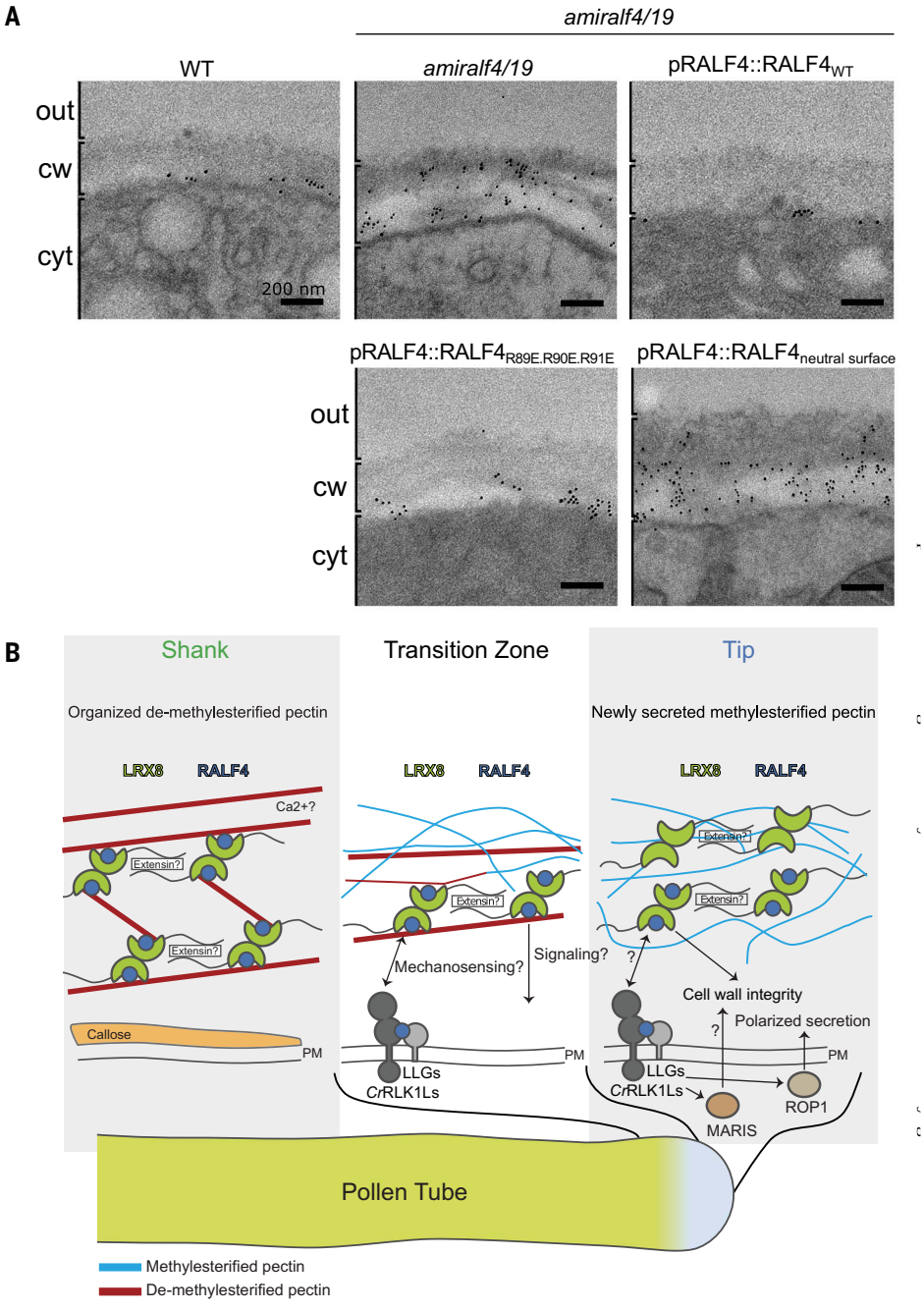


Fig. 4. LRX-RALF4-pectin patterning is required for cell wall assembly and integrity. (**A**) RALF4 surface variants display an aberrant cell wall architecture. Shown are TEM micrographs of pollen tube transverse sections (~10 μm) of Col-0 (WT), *amiralf4/19*, and complementation lines expressing *pRALF4::RALF4*$_{WT}$ *pRALF4::RALF4*$_{R89E.R90E.R91E}$, and *pRALF4::RALF4*$_{neutral\ surface}$. Black dots indicate immunogold labeling for callose. Out, outside; cw, cell wall; cyt, cytoplasm. (**B**) Proposed molecular model of the two RALF4-activated interacting protein families that maintain pollen tube cell wall integrity during growth. Methyl-esterified and de-methyl-esterified pectin are depicted as blue and red lines, respectively. At the tip, membrane-spanning *Cr*RLK1Ls and LLGs sense RALF4/19 to activate downstream MARIS and ROP1 proteins to regulate actin dynamics and promote the polarized secretion of cell wall components and regulatory proteins. At the transition zone, pectin shifts to the demethylesterified form through pectin methylesterase activity (*3*) and is then recognized by the LRX-RALF4 complex. The reticulated network of LRX-RALF4 compacts and chaperones the demethylesterified pectin into a criss-crossed filament network. This network contributes to the patterning, reinforcement, and expansion of the cell wall. At the shank, pectin de-methylesterifcation is complete, and the LRX-RALF4-pectin complex remains an integral part of the pollen cell wall. We speculate that the LRX-RALF4-pectin network acts as a load-bearing system during pollen tube expansion, compensating for its low cellulose content. The modularity of the LRX-extensin network, through the binding of different RALF peptides, would allow targeting and connecting different cell wall polymers or epitopes during cell wall assembly. How the RALF4 membrane and cell wall systems communicate, how these architectural changes in the cell wall are sensed by the pollen tube cell, and whether the peptide can navigate between the two systems will be interesting topics of future studies.

monomerized LRX8-RALF4 complex also induced pectin stiffening (negative ΔD) but showed a greatly reduced pectin-binding capacity (ΔF) relative to the dimeric protein (Fig. 3C), suggesting functional differences depending on the oligomeric state of the complex. The charge dependence of the pectin interaction was confirmed by the strongly reduced pectin binding and condensing capacity of LRX8-RALF4$_{neutral\ surface}$ (Fig. 3D).

These results demonstrate that LRX8-RALF4 has the capacity to recognize and condense demethylesterified pectins in a charge-dependent way. To investigate whether this capacity may direct the patterning of the pollen tube cell wall and affect cell expansion, we generated complementation lines of the *amiralf4/19* line (*9*) by expressing artificial microRNA (amiRNA)–resistant versions of RALF4$_{WT}$, RALF4$_{R89E.R90E.R91E}$, and RALF4$_{neutral\ surface}$ under the control of the native RALF4 promoter (fig. S11). The two RALF4 variants colocalized with LRX8 in the cell wall (figs. S5 and S12). Although RALF4$_{WT}$ fully restored the fertility and pollen tube growth of the *amiralf4/19* line, RALF4$_{R89E.R90E.R91E}$ and RALF4$_{neutral\ surface}$ conferred a moderate or severe reduction in vivo, respectively (Fig. 3, E to H) but nevertheless maintained their capacity to direct their growth toward the ovules (*31*). The growth defect of *amiralf4/19* pollen tubes reflected a higher (not lower) growth velocity and premature bursting rate. Both parameters were only partially complemented in the lines expressing RALF4$_{R89E.R90E.R91E}$ or RALF4$_{neutral\ surface}$ (fig. S13). In addition, in these lines, a large proportion of the surviving pollen tubes did not show continuous oscillatory growth (*32*) but rather a stop-and-go growth pattern often followed by bursting (Fig. 3, I and J). Similar growth phenotypes were observed in the pollen LRX mutants (*33*) and in a mutant (*ppme1*) with impaired pectin demethylesterification

(*5*, *34*) (fig. S14). These data suggest that the interaction of LRX-RALF4 with demethylesterified pectin is critical for pollen tube cell wall integrity and growth.

To investigate the impact of impaired pectin interaction on cell wall patterning, we studied the cell walls of surviving pollen tubes in *amiralf4/19*–complemented lines with $RALF4_{neutral\ surface}$. Immunolabeling experiments showed a wild-type spatial localization of pectin epitopes in *amiralf4/19 pRALF4::*$RALF4_{neutral\ surface}$ lines, with high-DM pectins (LM20 antibody) and no-DM pectins (LM19 antibody) at the pollen tube tip and shank, respectively (fig. S15) (*3*, *25*). dSTORM microscopy on *amiralf4/19 pRALF4:*$RALF4_{neutral\ surface}$ cell walls, labeled for low DM pectins (LM19), showed a denser array of punctae (Fig. 3, K and L) relative to the wild type, which lacked filamentous features (Fig. 2, G and H, and fig. S8). Transmission electron microscopy (TEM) on transverse sections through the shank of pollen tubes of both wild-type and *amiralf4/19* $RALF4_{WT}$ complementation lines revealed a compact cell wall with a thin inner secondary cell wall layer labeled with anti-callose antibodies. A thicker, unstructured pectic cell wall filled with patches of callose was observed in *amiralf4/19* pollen tubes (Fig. 4A and fig. S16). The expression of $RALF4_{neutral\ surface}$ did not restore a wild-type cell wall architecture, whereas $RALF4_{R89E.R90E.R91E}$ partially complemented the aberrant discontinuous structural pattern but maintained the thick callose layer and a fibrous, more loosely packed outer cell wall (Fig. 4A and fig. S16). Altogether, these results show that RALF4 is not only a signaling peptide but also a structural cell wall component that promotes the patterning of cell wall polysaccharides into a strong polymer assembly as part of the LRX8-RALF4-pectin complex.

Discussion

In this study, we have shown that cell wall polysaccharides are shaped into a reticulated network through a peptide-activated LRX8 complex that binds and condenses pectins, thus becoming an integral part of the cell wall and conferring integrity to the expandable cell wall during pollen tube growth (Fig. 4B and fig. S17). The pectin homogalacturonan, despite the considerable differences between the mechanical landscapes of plant cell walls and the animal extracellular matrix, can be considered the equivalent of animal glycosaminoglycans (GAGs). GAGs are also unbranched, charged polysaccharides that undergo compaction and rigidification through their interaction with GAG-binding cross-linking proteins (*35*, *36*). LRX8-RALF4 is a polysaccharide-binding protein complex that promotes mesoscale cell wall patterning. The particularity of this complex, as opposed to known GAG-binding proteins, lies in its modularity, with a polycationic peptide activating the LRX8 cell wall protein to bind polyanionic pectic polysaccharides. The LRX8-RALF4-pectin interaction appears to be essential for cell wall patterning and pollen tube growth in plants, as shown by the failure of the $RALF4_{R89E.R90E.R91E}$ and $RALF4_{neutral\ surface}$ variants to restore normal cell wall organization, growth, and fertility to *amiralf4/19* pollen tubes, where the severity of the complemented phenotype scales with the severity of the homogalacturonan-binding defect.

The RALF4 peptide is molded through its interaction with LRX8, thus exposing at the surface of the heterotetramer two patches of cationic residues with an affinity for polyanionic homogalacturonan. The charge-dependent compaction of the pectin matrix may combine polyelectrolyte complex formation (also observed for the unstructured free peptide) and homogalacturonan cross-linking by the LRX8-RALF4 heterotetramer. The latter is suggested by the reduced binding capacity of the pectin layer for monomerized LRX8-RALF4 (shown by QCM-D) relative to the native complex despite a comparable affinity for isolated OGs, as shown by calorimetry and TSAs. This property might be critical for LRX8-RALF4 bioactivity given the unability of the LRX8 monomer to complement the fertility phenotype of the *lrx* quadruple mutant (*10*).

Pollen tube growth involves a balancing act among cell wall assembly, consolidation, and turgor-driven wall expansion (*2*, *3*). We propose that the web-like LRX-RALF-pectin network constitutes a load-bearing component of the pollen tube cell wall, perhaps compensating for its low cellulose content (*3*, *37*). Disruption of this network causes the pollen tube to burst, in particular during the transition from the style into the transmitting tract (*18*). In this context, it is notable that LRX proteins comprise, in addition to a RALF-binding leucine-rich repeat domain, an extensin domain (*10*, *33*, *38*). Extensins are hydroxyproline-rich, periodically amphiphilic proteins, often with YXY motifs that can form extensin peroxidase-catalyzed, di-isodityrosine intermolecular cross-links with other extensins (*39*–*42*). At least one extensin (AtEXT3) has been shown to form a dendritic scaffold in vitro. Because of the polybasic nature of AtEXT3, it was hypothesized that this scaffold may structure the cell wall by creating pectin-extensin coacervates, which may be consolidated by oxidative cross-linking between extensins (*43*). The extensin domain of LRX8 contains cross-linking acceptor tyrosine residues, but it is much less charged than AtEXT3 (17 His residues versus 91 His or Lys residues in AtEXT3). We propose that LRX8 co-opts the positive charges of the polybasic RALF4 peptide to mediate specific pectin interactions. This modularity has two obvious advantages. First, it creates the possibility of targeting the LRX-extensin network to different cell wall polymers or polymer motifs by exchanging the RALF peptide. A large number of RALF peptides have the capacity to bind to the conserved binding pocket of LRX proteins, exposing distinct surface patches (*10*). This could contribute to the patterning of cell-type–specific cell walls or cell wall domains with different architectures, such as the reticulated homogalacturonan network shown in onion cell walls or the pectin nanofilaments in epidermal cells (*27*, *44*). The second advantage is that the modularity would allow rapid regulation of cell wall assembly by monitoring free RALF peptides, which may feed back to cellular processes that control actin dynamics, as well as secretion and modification of cell wall polymers, through the low-affinity LLG/BUPS/ANX receptor complexes (*12*, *16*, *45*–*52*). Future studies could investigate under which conditions free RALF peptides are released or transferred to instruct the *Cr*RLK1Ls and LRX complexes and how this informs the cell on the status of the cell wall assembly process during cell expansion and morphogenesis. In one possible scenario, pectin could connect LLG/*Cr*RLK1L to LRX-RALF complexes, thus forming a mechanosensing cell wall network (*45*). Supporting this scenario, a direct interaction between pectin and *Cr*RLK1Ls has been reported (*53*). This interaction is most likely weak, at least for the *Cr*RLK1Ls ANX1 and BUPS1, because it escaped detection by calorimetry or TSAs (*20*, *21*) when tested with different OGs or polygalacturonic acid (fig. S18) (*54*). Nevertheless, such a weak interaction might be suitable for supporting a role in mechanosensing, for example, by detecting the deformation of the LRX-RALF-pectin network. This study uncovers the unique dual signaling and structural role of RALF4 in shaping cell wall assembly during cell expansion and morphogenesis and may provide inspiration to uncover additional roles for protein-polysaccharide interactions in defining cell wall architecture and properties.

REFERENCES AND NOTES

1. M. A. Atmodjo, Z. Hao, D. Mohnen, *Annu. Rev. Plant Biol.* **64**, 747–779 (2013).
2. P. Fayant *et al.*, *Plant Cell* **22**, 2579–2593 (2010).
3. Y. Chebli, M. Kaneda, R. Zerzour, A. Geitmann, *Plant Physiol.* **160**, 1940–1955 (2012).
4. E. Parre, A. Geitmann, *Planta* **220**, 582–592 (2005).
5. N. Röckel, S. Wolf, B. Kost, T. Rausch, S. Greiner, *Plant J.* **53**, 133–143 (2008).
6. G. Pearce, D. S. Moura, J. Stratmann, C. A. Ryan Jr., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12843–12847 (2001).
7. M. Haruta, G. Sabat, K. Stecker, B. B. Minkoff, M. R. Sussman, *Science* **343**, 408–411 (2014).
8. E. Murphy *et al.*, *J. Exp. Bot.* **67**, 4863–4875 (2016).
9. M. A. Mecchia *et al.*, *Science* **358**, 1600–1603 (2017).
10. S. Moussu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7494–7503 (2020).
11. M. Gonneau *et al.*, *Curr. Biol.* **28**, 2452–2458.e4 (2018).
12. Z. Ge *et al.*, *Science* **358**, 1596–1600 (2017).
13. S. Schoenaers *et al.*, *Curr. Biol.* **28**, 722–732.e6 (2018).
14. M. Stegmann *et al.*, *Science* **355**, 287–289 (2017).
15. A. Herger *et al.*, *PLOS Genet.* **16**, e1008847 (2020).
16. Z. Ge *et al.*, *Curr. Biol.* **29**, 3256–3265.e5 (2019).

17. B. C. W. Crawford, G. Ditta, M. F. Yanofsky, *Curr. Biol.* **17**, 1101–1108 (2007).
18. I. Fobis-Loisy, Y. Jaillais, *Dev. Cell* **56**, 873–875 (2021).
19. A. Geitmann, *Sex. Plant Reprod.* **23**, 63–71 (2010).
20. C. J. Layton, H. W. Hellinga, *Protein Sci.* **20**, 1439–1450 (2011).
21. P. J. Sandoval, J. Santiago, *Plant Physiol.* **182**, 1697–1712 (2020).
22. M. Bosch, A. Y. Cheung, P. K. Hepler, *Plant Physiol.* **138**, 1334–1346 (2005).
23. E. Betzig *et al*., *Science* **313**, 1642–1645 (2006).
24. M. J. Rust, M. Bates, X. Zhuang, *Nat. Methods* **3**, 793–796 (2006).
25. Y. Verhertbruggen, S. E. Marcus, A. Haeger, J. J. Ordaz-Ortiz, J. P. Knox, *Carbohydr. Res.* **344**, 1858–1862 (2009).
26. K. Palacio-Lopez *et al*., *Front. Plant Sci.* **11**, 1032 (2020).
27. K. T. Haas, R. Wightman, E. M. Meyerowitz, A. Peaucelle, *Science* **367**, 1003–1007 (2020).
28. F. Dardelle *et al*., *Plant Physiol.* **153**, 1563–1576 (2010).
29. Y.-Q. Li *et al*., *Planta* **200**, 41–49 (1996).
30. M. C. Dixon, *J. Biomol. Tech.* **19**, 151–158 (2008).
31. S. Okuda *et al*., *Nature* **458**, 357–361 (2009).
32. J. A. Feijó *et al*., *BioEssays* **23**, 86–94 (2001).
33. T. N. Fabrice *et al*., *Plant Physiol.* **176**, 1981–1992 (2018).
34. G.-W. Tian, M.-H. Chen, A. Zaltsman, V. Citovsky, *Dev. Biol.* **294**, 83–91 (2006).
35. R. P. Richter, N. S. Baranova, A. J. Day, J. C. Kwok, *Curr. Opin. Struct. Biol.* **50**, 65–74 (2018).
36. J. J. Walkowiak, M. Ballauff, R. Zimmermann, U. Freudenberg, C. Werner, *Biomacromolecules* **21**, 4615–4625 (2020).
37. H. Schlupmann, A. Bacic, S. M. Read, *Plant Physiol.* **105**, 659–670 (1994).
38. C. Ringli, *Plant J.* **63**, 662–669 (2010).
39. J. D. Brady, I. H. Sadler, S. C. Fry, *Biochem. J.* **315**, 323–327 (1996).
40. M. A. Held *et al*., *J. Biol. Chem.* **279**, 55474–55482 (2004).
41. S. Moussu, G. Ingram, *Cell Surf.* **9**, 100094 (2023).
42. D. T. A. Lamport, M. J. Kieliszewski, Y. Chen, M. C. Cannon, *Plant Physiol.* **156**, 11–19 (2011).
43. M. C. Cannon *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2226–2231 (2008).
44. W. J. Nicolas *et al*., *Curr. Biol.* **32**, 2375–2389.e6 (2022).
45. X. Zhou *et al*., *Dev. Cell* **56**, 1030–1042.e6 (2021).
46. N. Luo *et al*., *Nat. Commun.* **8**, 1687 (2017).
47. A. Boisson-Dernier *et al*., *PLOS Biol.* **11**, e1001719 (2013).
48. L. Zhu *et al*., *Plant J.* **95**, 474–486 (2018).
49. Y. J. Lee, A. Szumlanski, E. Nielsen, Z. Yang, *J. Cell Biol.* **181**, 1155–1168 (2008).
50. S. T. McKenna *et al*., *Plant Cell* **21**, 3026–3040 (2009).
51. Y. Gu *et al*., *J. Cell Biol.* **169**, 127–138 (2005).
52. Y. Gu, V. Vernoud, Y. Fu, Z. Yang, *J. Exp. Bot.* **54**, 93–101 (2003).
53. W. Feng *et al*., *Curr. Biol.* **28**, 666–675.e5 (2018).
54. S. Moussu, S. Augustin, A.-O. Roman, C. Broyart, J. Santiago, *Acta Crystallogr. D Struct. Biol.* **74**, 671–680 (2018).

ACKNOWLEDGMENTS

We thank A. Peaucelle for insightful discussions during project development and D. Mazaud, P. le Baccon, and D. Choquet for support. dSTORM experiments were performed at the Curie Institute, Paris, and the Interdisciplinary Institute for Neurocience (IINS), Bordeaux. **Funding:** This work was supported by the University of Lausanne (J.S.); the European Research Council (WallWatchers grant 716358 to J.S. and STORMtheWALL grant 101041597 to K.T.H.); the Swiss National Science Foundation (grant 310030_204526 to J.S. and grant CR32I3_156724 to U.G.); French Bio-Imaging (grant 2411 to K.T.H.); Agence Nationale de Recherche (Homeowall grant to H.H.); Saclay Plant Sciences (grant ANR-17-EUR-0007 to H.H. and K.T.H.); Research Foundation Flanders (grants 1225120N and G013023N); the University of Antwerp (grants BOF-KP and BOF-DOCPRO4 to K.V.); and LASERLAB-EUROPE (grant 654148 to S.S.). **Author contributions:** S.M. and H.K.L. were responsible for conceptualization, methodology, investigation, visualization, and data analysis. J.S. and S.M. performed the structural analysis and designed the RALF4 mutants. S.M. generated all transgenic lines reported in the study. S.M. and H.K.L. performed the pollen tube growth assays, TSA and ITC assays, and immunostaining experiments; prepared samples for dSTORM, AiryScan, and TEM imaging; and prepared Figs. 1; 2, A and I; 3, E to J; and 4, A to D, and the associated supplementary information. K.T.H prepared Figs. 2, B to E, G, H, and J; Fig. 3, K and L; and figs. S8 and S9. K.T.H. acquired and analyzed dSTORM data and prepared the associated figures. C.B. prepared Fig. 1, E and J, and figs. S2A, S3, S4C, and S11. C.B. purified proteins, performed ITC assays, and prepared the corresponding figures. U.R. acquired and analyzed Airyscan data and prepared the associated figures (Fig. 2F and fig. S7). D.D.B. performed the TEM sectioning and immunodetection of callose and acquired the TEM data images. T.L. and E.B. prepared pectins and OGs with different DMs. G.S.-F. generated the first generation of in vivo RALF4 mutants. U.G. provided materials and supervised the initial work on the first generation of in vivo RALF4 mutants. S.S. participated in the optimization of pollen tube growth behavior. K.V. supervised the optimization of pollen tube growth behavior. E.H. provided dSTORM support. N.G. supervised the Airyscan data acquisition and analysis. B.C. provided QCM-D expertise and supported data analysis. H.H. and T.L. performed the QCM-D experiments and data analysis and prepared the associated figures. H.H. and T.L. prepared Fig. 3, A to D, and fig. S10. J.S. and H.H. were responsible for conceptualization, supervision, funding acquisition, and project administration. J.S. and H.H. wrote the original draft. All authors reviewed and edited the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Materials are available through the corresponding authors upon request. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse. This research was funded in whole or in part by the European Research Council (grants 716358 and 101041597), the Swiss National Science Foundation (grant 310030_204526), Agence Nationale de Recherche (grant "Homeowall"), Saclay Plant Sciences (grant ANR-17-EUR-0007), and the Swiss National Science Foundation (grant CR32I3_156724), all cOAlition S organizations. The author will make the Author Accepted Manuscript (AAM) version available under a CC BY public copyright license.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi4720
Materials and Methods
Figs. S1 to S18
References (*55*–*63*)
MDAR Reproducibility Checklist

Submitted 27 April 2023; accepted 22 September 2023
10.1126/science.adi4720

EPIGENETICS

Histone removal in sperm protects paternal chromosomes from premature division at fertilization

Raphaëlle Dubruille[1]*†, Marion Herbette[1]†‡, Maxime Revel[1]§, Béatrice Horard[1], Ching-Ho Chang[2], Benjamin Loppin[1]*

The global replacement of histones with protamines in sperm chromatin is widespread in animals, including insects, but its actual function remains enigmatic. We show that in the *Drosophila* paternal effect mutant *paternal loss* (*pal*), sperm chromatin retains germline histones H3 and H4 genome wide without impairing sperm viability. However, after fertilization, *pal* sperm chromosomes are targeted by the egg chromosomal passenger complex and engage into a catastrophic premature division in synchrony with female meiosis II. We show that *pal* encodes a rapidly evolving transition protein specifically required for the eviction of $(H3\text{-}H4)_2$ tetramers from spermatid DNA after the removal of H2A-H2B dimers. Our study thus reveals an unsuspected role of histone eviction from insect sperm chromatin: safeguarding the integrity of the male pronucleus during female meiosis.

Sperm chromatin is generally characterized by a high level of DNA compaction, which reduces nuclear volume and contributes to the shape and hydrodynamic properties of the sperm head (*1*, *2*). In many animal species, tight packaging of sperm DNA follows the replacement of nucleosomal histones with sperm nuclear basic proteins (SNBPs), such as the well-characterized mammalian protamines (*3*). This distinct, global chromatin remodeling process known as the histone-to-protamine transition occurs during spermiogenesis, the differentiation of post-meiotic spermatids (*1*, *2*, *4*). Although the histone-to-protamine transition is generally assumed to be essential for the formation of functional sperm, not all animals use SNBPs for sperm chromatin assembly. For example, whereas mammalian sperm DNA is mostly packaged with protamines, many other vertebrates maintain a full nucleosome-based sperm chromatin (*5*, *6*). Even within groups of related species (e.g., teleost fishes), sperm chromatin can vary between protamine based and histone based (*7*, *8*). These repeated transitions between nucleosomal and nucleoprotamine sperm chromatin (*3*, *9*) have left the function of histone replacement by SNBPs unclear.

Despite their enormous diversity and ancient evolutionary history, all major orders of insects consistently encode "histone-free" packaging of sperm DNA (*10*–*12*). As the best characterized example, the model species *Drosophila melanogaster* has ultracompact sperm chromatin entirely organized with protamine-like SNBPs, except for centromeres that retain the centromeric histone H3 (CenH3) (*1*, *4*, *13*). This indicates that a stringent selective requirement for SNBP-based sperm chromatin must exist in insects. In this work, we discovered that the replacement of histones with SNBPs is functionally linked to a postfertilization process through the characterization of a *Drosophila* paternal effect mutant named *paternal loss* (*pal*).

[1]Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS UMR5239, Université Claude Bernard Lyon 1, Lyon, France. [2]Division of Basic Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA.
*Corresponding author. Email: raphaelle.renard-dubruille@ens-lyon.fr (R.D.); benjamin.loppin@ens-lyon.fr (B.L.)
†These authors contributed equally to this work.
‡Present address: Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK.
§Present address: Department of Genetics, University of Geneva, Geneva, Switzerland.

Results

Paternal effect mutants represent a rare genetic resource to investigate sperm functions that are specifically required for the integration of paternal chromosomes in the zygote (*14*–*18*). The *pal* mutant, which was isolated more than fifty years ago, has been known to induce sporadic loss of paternal chromosomes at the onset of embryonic development (*14*, *19*, *20*), but the origin of this phenotype is unknown. In wild-type eggs fertilized by sperm from *pal* mutant males, loss of the paternal X chromosome, for example, occurs in about 3% of adult progeny and can lead to the development of bilateral gynandromorphs, with half male (X0) and half female (XX) tissues (*20*) (Fig. 1A). Similarly, losses of the Y chromosome and the small chromosome 4, but not one of the larger, essential chromosomes 2 or 3, were also detected in the progeny of *pal* males (*14*, *20*). To investigate the cytological bases of these early chromosome losses, we collected wild-type eggs fertilized by sperm from *pal* mutant males (hereafter referred to as *pal* eggs). In *Drosophila*, after fertilization, the male and female pronuclei migrate to a central region of the egg where they appose without fusing their nuclear envelopes (Fig. 1B). We observed frequent (~50 to 60%; table S1) fragmentation of the m pronucleus at the apposition stage in *pal* eggs (Fig. 1C). As expected, this phenotype was followed by the sporadic loss or fragmentation of paternal chromosomes during the first and second embryonic mitoses (Fig. 1, D and E), which was in agreement with Baker's original prediction based on adult mosaic analyses (*20*). In addition, we observed that about half of *pal* eggs did not hatch (table S2), suggesting that most of these early chromosomal defects cause embryonic arrest. Taken together, these observations indicate that loss of paternal chromosomes in *pal* eggs is a secondary consequence of a much more penetrant phenotype that affects the integrity of the male pronucleus.

Paternal chromosomes in *pal* eggs are aberrantly targeted by the CPC

We investigated the cause of male pronuclear fragmentation in *pal* eggs. In contrast to most vertebrates in which fertilization occurs in oocytes arrested in metaphase II of female meiosis, insect fertilization occurs when female meiosis resumes from metaphase I arrest (*21*–*26*). In *Drosophila*, the earliest stage of fertilized eggs practically accessible to cytology is metaphase II, which is recognized by the presence of the female meiotic spindles organized in tandem (Fig. 1B). At this stage, in eggs fertilized by wild-type sperm (referred to as wild-type eggs), the sperm nucleus has already replaced its SNBPs with histones and typically appears as a small, roundish nucleus. We found that in *pal* eggs, the male nucleus was irregular in shape and almost systematically surrounded by a spindle of microtubules (96.4%, $n = 28$; Fig. 1F). Such ectopic, paternal spindles were not observed in wild-type eggs (0%, $n = 36$). Like female meiotic spindles, the paternal spindle in *pal* eggs was anastral (Fig. 1F) and sperm-derived centrioles were normally found in the sperm aster, a large aster of microtubules that captures the female pronucleus (fig. S1). The paternal pseudodivision progressed in synchrony with female meiosis II and the paternal chromatids derived from the haploid sperm nucleus eventually stretched between the spindle poles (Fig. 1F). This suggests that paternal chromosomes in *pal* eggs follow the same cytoplasmic cues that control female meiosis progression. We thus investigated whether the pseudodivision of *pal* paternal chromosomes is triggered by the chromosomal passenger complex (CPC), which controls meiotic spindle formation in *Drosophila* oocytes by enabling kinetochore assembly and microtubule recruitment (*27*). A key subunit of the CPC is the Aurora B kinase, which phosphorylates histone H3 on serine 10 (H3S10ph) (*28*). In wild-type eggs in meiosis II, H3S10ph was exclusively found on maternal chromosomes ($n = 21$; Fig. 1F). By contrast, in *pal* eggs, this histone mark

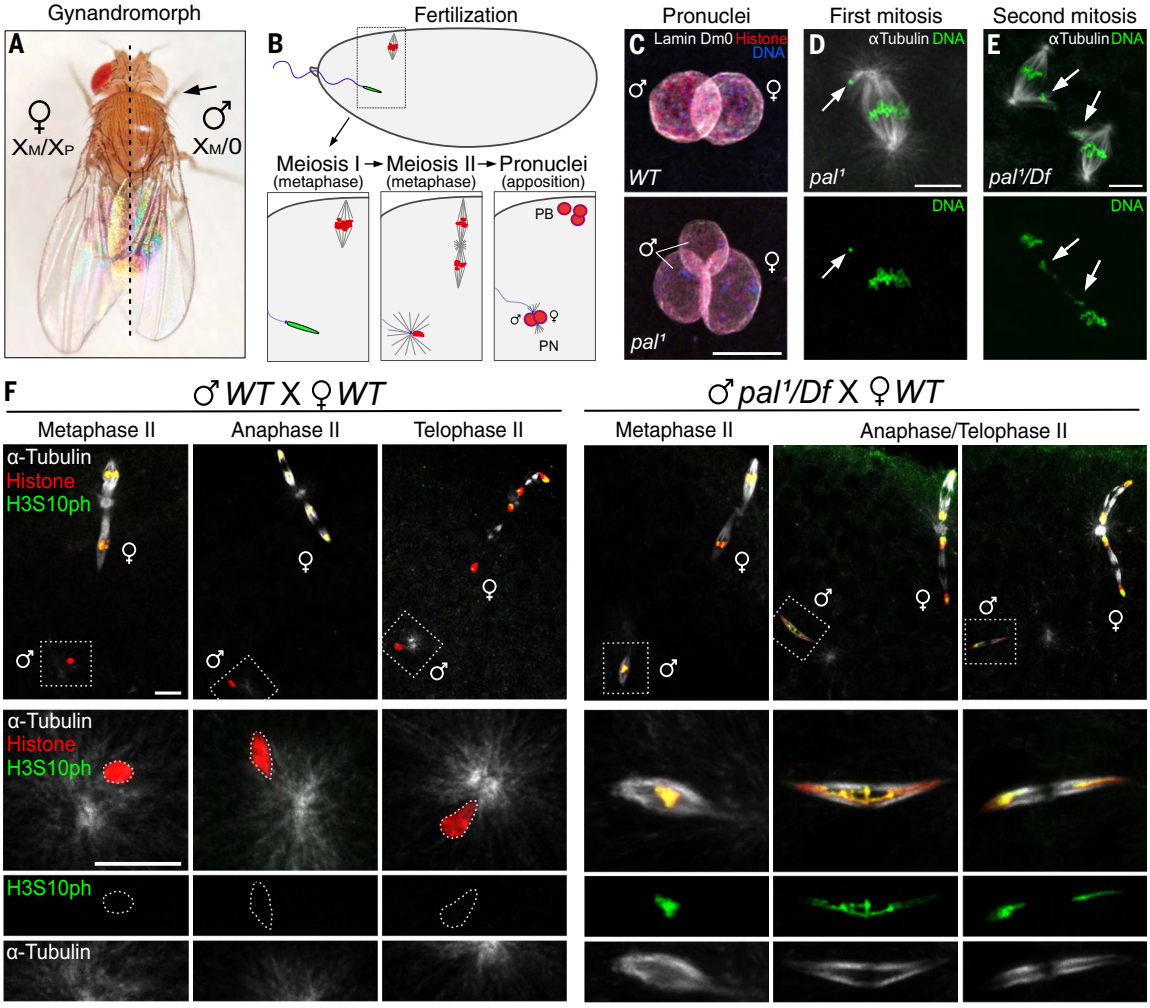


Fig. 1. *pal* induces premature condensation of paternal chromosomes at fertilization. (**A**) A gynandromorph obtained from a cross between *pal*^{1} males and *w*^{1118} females. Cells in the right half (male) are missing the *white+* (red eye color) paternal X chromosome (XP) and have only the *white* maternal X chromosome (XM). Note the shorter male wing and presence of sex combs on first leg (arrow). (**B**) Schematic of fertilization and zygote formation in *Drosophila*. Sperm chromatin (green), nucleosomal chromatin (red), microtubules (black), sperm tail, and centrioles (blue). PB, polar bodies; PN, pronuclei. (**C**) (Top) Apposed male and female pronuclei in a control egg. (Bottom) A fragmented male pronucleus in an egg fertilized by *pal*^{1} sperm. WT, wild type. (**D**) A *pal*^{1} embryo during first mitosis with one small chromosome positioned outside the main spindle (arrow). (**E**) Second mitosis in a *pal*^{1}*/Df(2L)ED690* embryo with genetic material scattered between the two spindles (arrows). (**F**) (Left) Control eggs at the indicated phase of female meiosis II. The male pronucleus (inset) remains associated with the sperm aster. (Right) In *pal*^{1}*/Df(2L)Exel6024* eggs at metaphase of female meiosis II, an ectopic spindle is formed around the male pronucleus, which aberrantly stains for H3S10ph. In anaphase and telophase II, the paternal chromosomes are stretched between the spindle poles. All scale bars, 10 μm.

decorated both maternal and paternal chromosomes (100%, n = 15; Fig. 1F). Consistently, INCENP, another CPC subunit (*27*), also aberrantly localized on paternal chromosomes and the paternal spindle in *pal* eggs (fig. S1). To demonstrate that the CPC was responsible for the cytological defects affecting the male pronucleus in *pal* eggs, we crossed *pal* males with females expressing a small hairpin RNA targeting *aurora B* (*aurB*) in the germline. As expected, maternal knockdown (KD) of *aurB* abolished phosphorylation of H3S10 on female chromosomes and prevented meiosis progression (fig. S2). In *aurB* KD eggs fertilized with *pal* sperm, we additionally observed a complete suppression of the *pal* pronuclear phenotype (fig. S2). We conclude that in *pal* eggs, the CPC instructs the paternal genome to engage in a pseudomeiotic division, leading to frequent pronuclear fragmentation and occasional losses of paternal chromosomes.

Histones H3 and H4 are aberrantly retained in *pal* sperm chromatin

Because *pal* is a paternal effect mutant, we investigated whether these egg phenotypes may result from an altered sperm chromatin organization. In *Drosophila*, the histone-to-protamine transition occurs at midspermiogenesis and noncentromeric histones are replaced with protamine-like SNBPs before the individualization of spermatids (*1*). *Drosophila* comprises a minimal set of nucleosome core histones, including canonical (or replicative) histones H2A, H2B, H3.2, and H4, and histone variants H2Av, H3.3, and Cid/CenH3 (*29*). We found that transgenic red fluorescent protein (RFP)–tagged H2A, H2Av, and H2B expressed in the male germline were completely removed from spermatid nuclei at the histone-to-protamine transition in control and *pal* testes (Fig. 2, A and B, and fig. S3, A and B). By contrast, although green fluorescent protein (GFP)–tagged H3.2 and H3.3 were also removed in control males, both transgenic H3 histones were aberrantly retained in *pal* mutants, marking all sperm nuclei (Fig. 2C and fig. S3C). Our immunofluorescence analyses also revealed

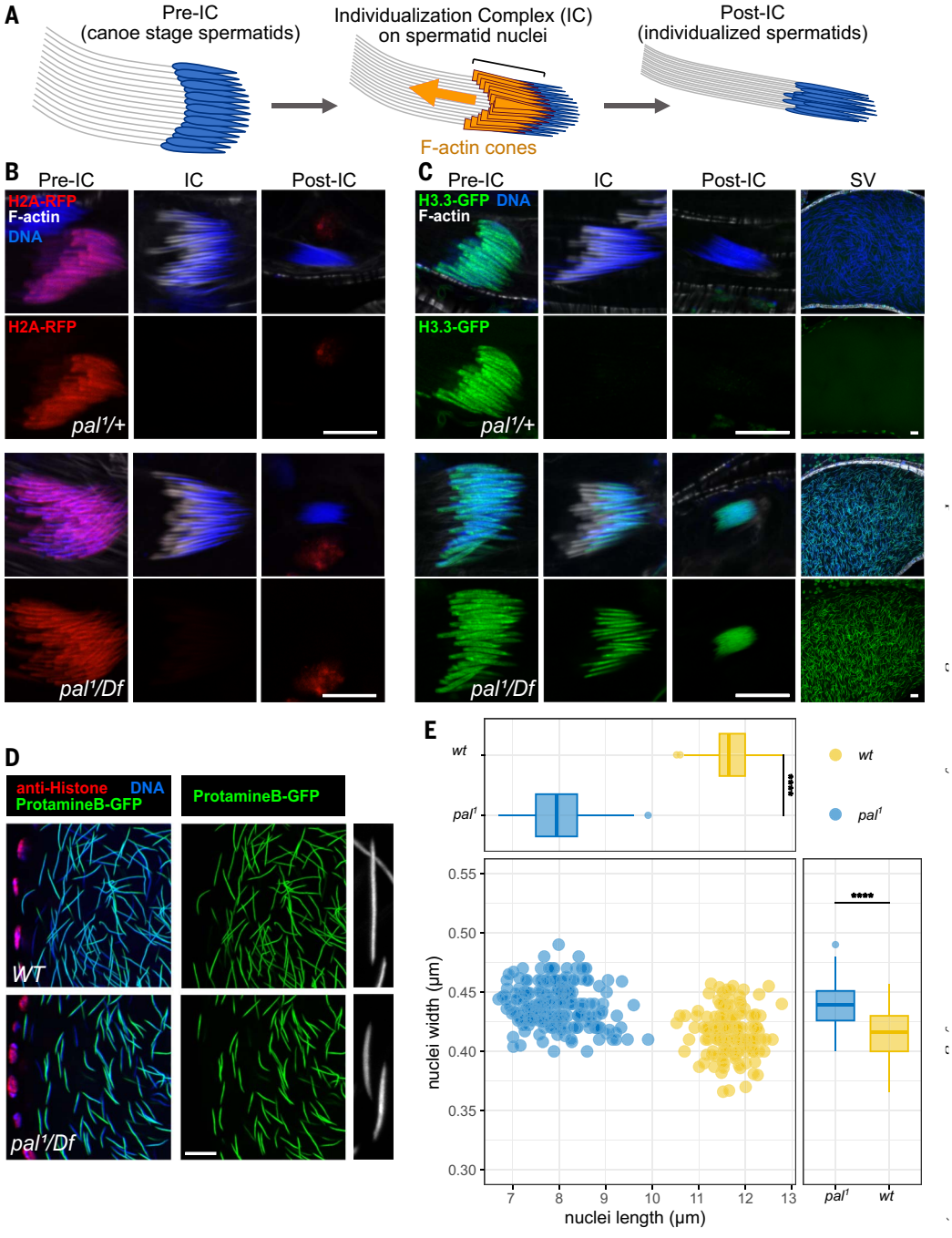


Fig. 2. H3 and H4 are retained in *pal* sperm chromatin. (**A**) Schematic illustrating the spermiogenesis stages shown in (B) and (C). F-actin cones forming the individualization complex (IC) progress from the anterior end (right) to the posterior end (left) of spermatids. The presence of IC (orange) on spermatid nuclei (blue) indicates the end of the histone-to-protamine transition. (**B**) Confocal images of spermatid nuclei at the indicated stage. In both control (pal^1/+) and pal^1/*Df(2L) Exel6024* testes, H2A-RFP is eliminated at the histone-to-protamine transition, before the formation of individualization cones (F-actin). (**C**) H3.3-GFP is eliminated from control spermatids but is aberrantly retained in pal^1/*Df(2L) Exel6024* spermatids and mature spermatozoa in seminal vesicles (SV). (**D**) *pal* sperm nuclei incorporate ProtamineB-GFP (ProtB-GFP). Mutant sperm nuclei have aberrant morphology of compared with control sperm nuclei (close-up, right). Scale bars, 10 µm. (**E**) Measurements of sperm nuclear length and width in wild-type (WT) (n = 136) and pal^1 (n = 194) seminal vesicles. t test, ****P < 0.0001.

the persistence of acetylated H4 (H4ac) beyond the histone-to-protamine transition in *pal* testes (fig. S3D), thus suggesting that histones are retained as $(H3\text{-}H4)_2$ tetramers on sperm DNA (tetrasomes) following the partial disassembly of nucleosomes. We observed that GFP-Cid, the *Drosophila* centromeric H3 that is retained in mature sperm (*13*), was normally distributed as four foci in *pal* sperm nuclei (fig. S3E). These results imply that, during normal *Drosophila* spermiogenesis, the removal of H2A-H2B dimers from spermatid nucleosomes is functionally uncoupled from the final elimination of $(H3\text{-}H4)_2$ tetramers. To directly examine the dynamics of histone removal in transitioning spermatid nuclei, we simultaneously imaged H2A-RFP and H3.3-GFP and measured relative fluorescence intensities in spermatid cysts at different stages. We found that the disassembly of spermatid nucleosomes is a sequential process, with H2A-H2B dimers being eliminated first, followed by histones H3 and H4 (fig. S4). Notably, global retention of H3 and H4 in *pal* spermatozoa did not preclude loading of the major *Drosophila* SNBPs Protamine B (ProtB) and Mst77F (Fig. 2D and fig. S3A). However, western blot quantification of ProtB-GFP suggested that SNBPs are less abundant in *pal* sperm than in control sperm (fig. S5). Although *pal* sperm nuclei appeared regular and homogeneous in shape, they were on average 25% shorter and significantly wider than wild-type sperm nuclei (Fig. 2, D and E). This difference had no apparent impact on the motility or fertilization potential of *pal* gametes (Movie S1 and fig. S6), revealing the remarkable plasticity of *Drosophila* spermatozoa with respect to their chromatin composition and nuclear morphology. We conclude that the *pal* gene is specifically required for the elimination of H3 and H4 at the histone-to-protamine transition.

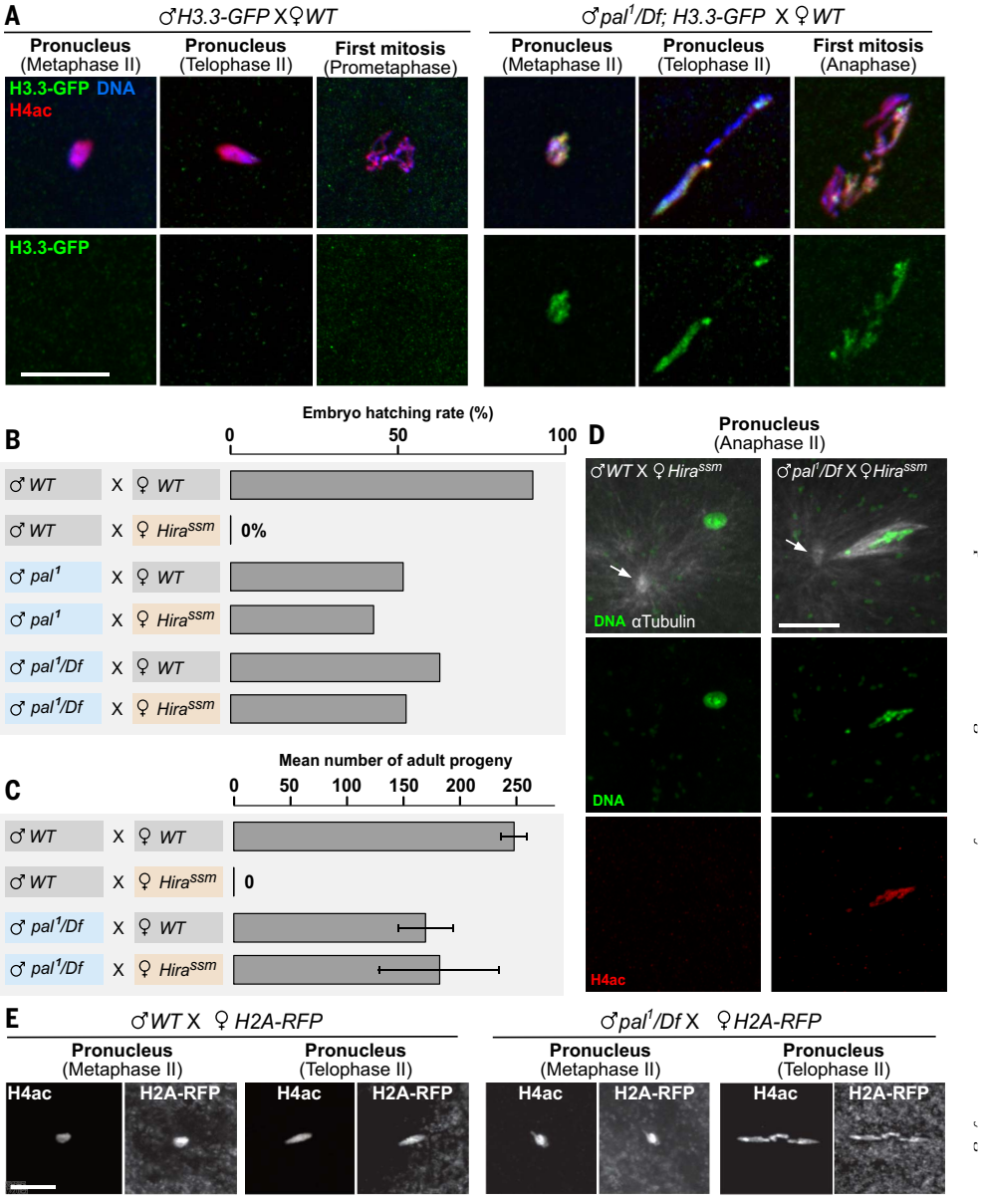


Fig. 3. *pal* sperm rescues *Hira* maternal effect embryo lethality. (**A**) H3.3-GFP is paternally transmitted by *pal*^{1}*/Df(2L)Exel6024* males and is detected on paternal chromatin during the first mitosis. WT, wild type. (**B**) Percentage of hatched embryos obtained from the indicated crosses. *Df* represents *Df(2L) Exel6024*. (**C**) Mean number (±SD) of F1 adult progeny produced by the indicated crosses. (**D**) In *Hira*^{ssm} control eggs (left), the male pronucleus is round and does not stain for H4ac. In *Hira*^{ssm} eggs fertilized with sperm from *pal*^{1}*/Df(2L)Exel6024* males (right), a spindle forms around paternal chromosomes, which stain for H4ac. Arrows indicate the center of the sperm aster. Both eggs are infected with the natural endosymbiont *Wolbachia*, which is visible as faint dots in the green channel. (**E**) Maternal H2A-RFP is deposited in paternal chromatin during decondensation of the male pronucleus in both WT and *pal*^{1}*/Df(2L)Exel6024* eggs. All scale bars, 10 μm.

pal sperm rescues the viability of *Hira* mutant eggs

We then sought to evaluate the actual contribution of *pal* sperm histones to paternal nucleosome reassembly after fertilization. Crossing H3.3-GFP–expressing *pal* males with wild-type females revealed that H3.3-GFP specifically decorated paternal chromatin in the male pronucleus and throughout the first zygotic division (100%, n = 18; Fig. 3A). Thus, paternally transmitted histones in *pal* sperm directly contribute to the formation of paternal nucleosomes in the egg. We then wondered whether maternal H3 and H4 histones stored in the egg are nevertheless required for male pronuclear formation in *pal* eggs. In both vertebrates and invertebrates, genome-wide histone deposition on paternal DNA following SNBP removal is initiated by the conserved HIRA histone chaperone complex (*30*). Specifically, maternal HIRA assembles $(H3.3\text{-}H4)_2$ tetramers in the male pronucleus in a replication-independent manner. In the absence of functional HIRA, such as in the *Drosophila* maternal effect mutant *sésame* (*Hira*^{ssm}), fertilized eggs fail to initiate paternal nucleosome assembly and develop only with maternal chromosomes, which leads to haploid development and full embryonic lethality (*31*, *32*). We observed that crosses between *Hira*^{ssm} females and *pal* males were fertile, with embryo hatching at rates on par with that of control crosses involving *pal* males and wild-type females (Fig. 3B and table S2). In addition, both crosses produced a similar number of adult progenies (Fig. 3C), and *pal* males were also able to restore the fertility of females with germline knockdown of *Hira* (table S2). At the cytological level, instead of the typical *Hira*^{ssm} phenotype showing a round, inert male pronucleus, *Hira*^{ssm} eggs fertilized by *pal* sperm showed only a *pal* phenotype, demonstrating that *pal* is epistatic to *Hira*^{ssm} (Fig. 3D). This suggests that, in *pal* sperm chromatin, H3 and H4 are present as tetrasomes throughout the paternal genome, thus bypassing the need for HIRA during male pronuclear formation. To determine the timing of paternal nucleosome reconstitution in *pal* eggs, we crossed *pal* mutant males with transgenic females expressing H2A-RFP. Maternal H2A-RFP decorated the male pronucleus of *pal* eggs as early as meiosis II (100%, n = 13) in a way indistinguishable from that of control eggs

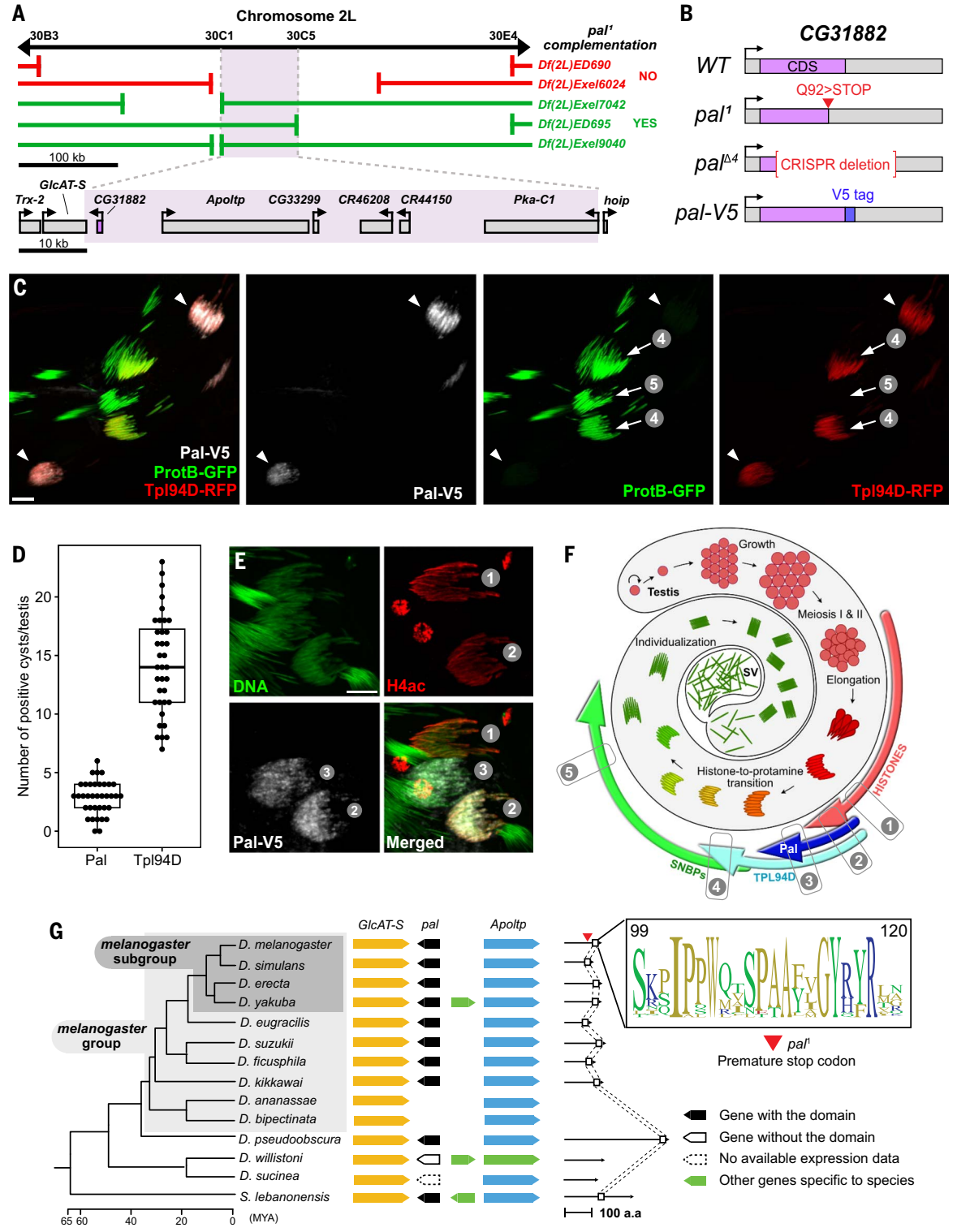


Fig. 4. Pal is a fast-evolving transition protein. (**A**) Deficiency mapping of *pal*[1]. The *CG31882/pal* genetic region is shown in purple. (**B**) Schematic of *CG31882/pal* alleles and *pal-V5* rescue transgene. (**C**) Pal-V5 is detected in spermatid nuclei that are positive for Tpl94D-RFP but negative for ProtB-GFP (arrowheads). The numbers in gray circles correspond to the stages defined in (F). (**D**) Quantification of Pal-V5– and Tpl94D-RFP–positive cysts per testis indicates a shorter residency time of Pal in spermatid nuclei. (**E**) Pal-V5 is detected shortly before and after H4ac disappearance. (**F**) Schematic testis that shows the progression of the histone-to-protamine transition. The stages (one to five) are defined by the presence or absence of indicated chromosomal proteins [see also (C) and (E)]. (**G**) (Left) Representation of *pal* homologs and their neighboring genes, *GlcAT-S* and *Apoltp*, in 13 *Drosophila* species and an outgroup species, *Scaptodrosophila lebanonensis* (subfamily Drosophilinae), whose phylogenetic relationships and divergence times are indicated on the left. There is an apparent absence of a *pal* homolog in *D. ananassae* and *D. bipectinata*. (Right) Pal proteins share a 21-aa conserved C-terminal domain, with relative size of letters indicating aa frequency at each position. Species (nonexhaustive) of the *D. melanogaster* group and subgroup are shown (gray backgrounds).

(100%, n = 11) (Fig. 3E). This result confirms that in *pal* eggs, H2A-H2B dimers stored in the egg cytoplasm are deposited on sperm tetrasomes in a replication-independent manner.

We conclude that the persistence of H3-H4 in *pal* sperm establishes an aberrant epigenetic identity of paternal chromosomes at fertilization, ultimately leading to their incorrect recognition by the CPC as maternal chromosomes during the second female meiotic division.

pal encodes a fast-evolving transition protein

To identify the *pal* gene, we performed a classical deficiency mapping of the original pal^1 allele (*19*, *20*). We narrowed down the *pal* locus to a 76-kb interval in region 30C of chromosome 2L (Fig. 4A and table S1; see the supplementary text). Across this region, whole-genome sequencing of pal^1 flies identified a nonsense point mutation in *CG31882*, a small, intronless gene specifically expressed in testes and predicted to encode a 121–amino acid (aa) basic protein (Fig. 4B and fig. S7). The premature stop codon in pal^1 leads to 30-aa C-terminal truncation of CG31882 protein. To validate that *CG31882* is indeed *pal*, we generated a deletion that removes most of *CG31882* coding region with CRISPR-Cas9. The deletion, named $pal^{\Delta 4}$, recapitulated the phenotypes of *pal* eggs (Fig. 4B and table S2). By contrast, we found that a previously reported insertion-deletion allele, $CG31882^{RT4}$ (*33*), which likely causes changes in the N terminus of the protein, did not impair *pal* function (fig. S7; see supplementary text). We also generated transgenes expressing wild-type Pal or Pal tagged with V5 (Pal-V5) under the endogenous regulatory regions of *pal* (Fig. 4B), which rescued the cytological defects of *pal* eggs (table S1). Both transgenes also restored full embryo lethality in crosses with $Hira^{ssm}$ females (table S2), thus confirming that the *CG31882* is the pal gene and demonstrating that Pal-V5 is functional.

At the cytological level, in testes, we observed that Pal-V5 transiently accumulates in spermatid nuclei undergoing the histone-to-protamine transition (Fig. 4, C and D). Pal-V5 first appeared in spermatids that had begun to incorporate the transition protein Tpl94D (*10*) and were still positive for H4ac (Fig. 4, C and E). Pal-V5 was still detectable shortly after H4ac disappearance but vanished before that of Tpl94D (Fig. 4, D to E). We thus conclude that Pal is a transition protein required for histone H3 and H4 eviction before the final deposition of SNBPs (Fig. 4F).

Previous analyses had suggested that *CG31882* is an evolutionarily young, testis-specific gene limited to species of the *D. melanogaster* subgroup, identified through sequence homology searches (*33*–*35*). However, our examination of syntenic regions flanked by *GlcAT-S* and *Apoltp* revealed orthologs of this gene in most surveyed *melanogaster* group species as well as in other Drosophilinae subfamily species (Fig. 4G and fig. S8A). RNA sequencing analyses further confirmed their testis- or male-specific expression (table S3). Comparison of the different *pal* orthologs reveals that its protein evolutionary rate [the nonsynonymous substitution rates (dN) normalized by synonymous substitutions rates (dS)] in the *D. melanogaster* subgroup is higher than that of most genes, including known SNBPs (fig. S8B) (*36*, *37*). Although we did not detect any positively selective sites using the site model in PAML (P = 0.59), the highly divergent Pal proteins share a conserved 20-aa motif at their C terminus (Fig. 4G). This motif, which is truncated in the pal^1 allele, is thus likely of critical importance for Pal function. In some *melanogaster* group species, such as *D. ananassae*, *pal* orthologs are apparently absent despite their conservation in more divergent species (Fig. 4G). Our findings are reminiscent of another study in *Drosophila* in which it was found that young, rapidly evolving SNBP genes are more likely to encode important functions than older, evolutionarily constrained genes (*37*). Thus, even though the histone eviction function is critical, the genetic means by which this function is mediated may vary across *Drosophila* and other insect species.

Discussion

Paternal effect mutants represent invaluable genetic tools to investigate the influence of spermiogenesis and sperm chromatin composition on the fate of paternal chromosomes at fertilization. In *Drosophila*, the molecular characterization of *K81* (*15*, *38*) and *deadbeat* (*17*) has, for example, revealed that these paternal gene products establish an epigenetic protection of sperm telomeres in the egg. Similarly, Heterochromatin Protein 1E exerts a transgenerational effect on the integrity of paternal chromosomes during the first mitosis (*16*). The elucidation of the *pal* phenotype now sheds light on the role of sperm histone elimination as an adaptation to constraints exerted by the egg cytoplasm on the fertilizing sperm nucleus.

Our characterization of *pal* first reveals the extraordinary plasticity of insect sperm chromatin composition. Although *Drosophila* spermiogenesis is sensitive to the loss of particular SNBPs (*39*–*41*), the global retention of H3 and H4 histones observed in *pal* mutants has no detectable impact on sperm differentiation or function despite substantial changes in nuclear morphology. It suggests that major diversification of sperm chromatin composition in animals could result from minimal changes in genes involved in the histone-to-protamine transition.

In *pal* sperm, we hypothesize that $(H3\text{-}H4)_2$ tetramers are distributed throughout the genome, reflecting the original position of nucleosomes before the removal of H2A and H2B at the histone-to-protamine transition. These tetrasomes must then coexist with SNBPs that are subsequently deposited, although less abundantly than in normal sperm chromatin. Incidentally, this study also establishes that histone elimination during *Drosophila* spermiogenesis is a sequential process, with the removal of H2A-H2B and H3-H4 being temporally and functionally distinct. Notably, during mouse spermiogenesis, the replacement of H2A and H2B with testis-specific histone variants prepares the final replacement of nucleosomes with protamines (*42*–*44*), whereas in *Xenopus*, only H2A-H2B dimers are partially replaced with SNBPs, leading to the retention of H3 and H4 in sperm (*45*). The sequential elimination of histones in spermatids could thus be widespread in animals.

The recurrent packaging of sperm DNA with nonhistone proteins is a longstanding mystery in chromatin biology and evolution. Main hypotheses on this specialization include sperm nuclear compaction, transcriptional shutdown in spermatids, and sperm DNA protection against damages (*5*, *44*, *46*). Our work now suggests that the timing of egg fertilization with respect to female meiosis progression is a previously unexplored, key determinant of sperm chromatin evolution in animals. In insects, early fertilization (metaphase of meiosis I) inevitably exposes the decondensing sperm nucleus to egg cytoplasmic cues that initiate the second meiotic division of maternal chromosomes. We propose that the global and invariable replacement of histones with SNBPs in insects establishes an epigenetic identity of paternal chromosomes that prevents their deleterious interaction with the egg CPC. In vertebrate eggs, the relatively late oocyte fertilization timing (metaphase of meiosis II) may offer a more permissive environment to the fertilizing sperm nucleus during its transformation into a pronucleus. In support of this possibility, premature chromosome condensation of sperm chromosomes, a defect highly reminiscent of the *pal* phenotype, is frequently observed when in vitro fertilization or intracytoplasmic sperm injection is performed with immature mammalian oocytes (*47*, *48*). The temporal uncoupling of female meiosis II initiation and male pronuclear formation in vertebrates could have thus favored the extreme diversification of sperm chromatin types, including histone-based sperm chromatin. Still, other animals with an early type of fertilization, such as the model nematode *Caenorhabditis elegans*, maintain histone-based sperm chromatin (*49*), which suggests the evolution of additional, yet unknown mechanisms to protect the paternal genome during the formation of the diploid zygote.

REFERENCES AND NOTES

1. C. Rathke, W. M. Baarends, S. Awe, R. Renkawitz-Pohl, *Biochim. Biophys. Acta* **1839**, 155–168 (2014).
2. L. Moritz, S. S. Hammoud, *Front. Endocrinol. (Lausanne)* **13**, 895502 (2022).
3. J. D. Lewis, Y. Song, M. E. de Jong, S. M. Bagha, J. Ausió, *Chromosoma* **111**, 473–482 (2003).
4. S.-L. Hao, F.-D. Ni, W.-X. Yang, *Gene* **706**, 201–210 (2019).
5. R. Balhorn, *Genome Biol.* **8**, 227 (2007).
6. J. Ausió, R. González-Romero, C. L. Woodcock, *J. Struct. Biol.* **188**, 142–155 (2014).
7. N. Saperas, E. Ribes, C. Buesa, F. García-Hegart, M. Chiva, *J. Exp. Zool.* **265**, 185–194 (1993).
8. S.-F. Wu, H. Zhang, B. R. Cairns, *Genome Res.* **21**, 578–589 (2011).
9. J. M. Eirín-López, J. Ausió, *BioEssays* **31**, 1062–1070 (2009).
10. C. Rathke *et al*., *J. Cell Sci.* **120**, 1689–1700 (2007).
11. P. M. Ferree *et al*., *Sci. Rep.* **9**, 12194 (2019).
12. G. A. Orsi *et al*., *Nat. Commun.* **14**, 4187 (2023).
13. N. Raychaudhuri *et al*., *PLOS Biol.* **10**, e1001434 (2012).
14. K. R. Fitch, G. K. Yasuda, K. N. Owens, B. T. Wakimoto, *Curr. Top. Dev. Biol.* **38**, 1–34 (1998).
15. R. Dubruille *et al*., *Curr. Biol.* **20**, 2090–2099 (2010).
16. M. T. Levine, H. M. Vander Wende, H. S. Malik, *eLife* **4**, e07378 (2015).
17. T. Yamaki, G. K. Yasuda, B. T. Wakimoto, *Genetics* **203**, 799–816 (2016).
18. A. M. Binder, B. T. Wakimoto, C. Davis, J. Chmielewski, J. E. Tomkiel Dean, *Res. J. Dev. Biol.* **4**, 1 (2017).
19. L. Sandler, *Drosoph. Inf. Serv.* **47**, 68 (1971).
20. B. S. Baker, *Genetics* **80**, 267–296 (1975).
21. J. R. Von Stetina, T. L. Orr-Weaver, *Cold Spring Harb. Perspect. Biol.* **3**, a005553–a005553 (2011).
22. B. Loppin, R. Dubruille, B. Horard, *Open Biol.* **5**, 150076 (2015).
23. M. G. Riparbelli, R. Stouthamer, R. Dallai, G. Callaini, *Dev. Biol.* **195**, 89–99 (1998).
24. M. Sato, H. Tanaka-Sato, *J. Morphol.* **254**, 266–271 (2002).
25. H. Sakai, T. Yokoyama, S. Tomita, *J. Insect Physiol.* **139**, 104386 (2022).
26. O. Marescalchi, C. Zauli, V. Scali, *Mol. Reprod. Dev.* **63**, 89–95 (2002).
27. L.-I. Wang *et al*., *J. Cell Biol.* **220**, e202006018 (2021).
28. R. R. Adams, H. Maiato, W. C. Earnshaw, M. Carmena, *J. Cell Biol.* **153**, 865–880 (2001).
29. B. Horard, B. Loppin, *Chromosoma* **124**, 163–175 (2015).
30. B. Loppin, F. Berger, *Annu. Rev. Genet.* **54**, 121–149 (2020).
31. B. Loppin, M. Docquier, F. Bonneton, P. Couble, *Dev. Biol.* **222**, 392–404 (2000).
32. B. Loppin *et al*., *Nature* **437**, 1386–1390 (2005).
33. S. Kondo *et al*., *Genes Dev.* **31**, 1841–1846 (2017).
34. S. Chen, Y. E. Zhang, M. Long, *Science* **330**, 1682–1685 (2010).
35. E. L. Rivard *et al*., *PLOS Genet.* **17**, e1009787 (2021).
36. Drosophila 12 Genomes Consortium, *Nature* **450**, 203–218 (2007).
37. C.-H. Chang, I. Mejia Natividad, H. S. Malik, *eLife* **12**, e85249 (2023).
38. G. Gao, Y. Cheng, N. Wesolowska, Y. S. Rong, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4932–4937 (2011).
39. S. Kimura, B. Loppin, *Open Biol.* **6**, 160207 (2016).
40. S. Tirmarche *et al*., *G3* **4**, 2241–2245 (2014).
41. Z. Eren-Ghiani, C. Rathke, I. Theofel, R. Renkawitz-Pohl, *Cell Rep.* **13**, 2327–2335 (2015).
42. E. Montellier *et al*., *Genes Dev.* **27**, 1680–1692 (2013).
43. S. Barral *et al*., *Mol. Cell* **66**, 89–101.e8 (2017).
44. L. Arévalo, G. Esther Merges, S. Schneider, H. Schorle, *Reproduction* **164**, R57–R74 (2022).
45. M. Mann, M. S. Risley, R. A. Eckhardt, H. E. Kasinsky, *J. Exp. Zool.* **222**, 173–186 (1982).
46. R. Oliva, G. H. Dixon, *Prog. Nucleic Acid Res. Mol. Biol.* **40**, 25–94 (1991).
47. Y. Yamauchi, J. M. Riel, M. A. Ward, *J. Androl.* **33**, 229–238 (2012).
48. H. Schmiady, H. Kentenich, *Hum. Reprod.* **4**, 689–695 (1989).
49. S. L. Ooi, J. R. Priess, S. Henikoff, *PLOS Genet.* **2**, e97 (2006).

ACKNOWLEDGMENTS

We thank H. S. Malik, M. Francesconi, and M. Delattre for their helpful comments and suggestions on the manuscript. We are grateful to E. C. Lai, K. McKim, and the Bloomington Drosophila Stock Center for antibodies and fly stocks. We also thank S. Suchet for the initial observation that *pal* males rescue the fertility of *Hira*^{ssm} females, M. Caron for plasmid constructions, and J. Brocard for his help with movie recording. We acknowledge the contribution of the Lyon SFR Biosciences (UAR3444/CNRS, US8/INSERM, ENS de Lyon, UCBL) imaging facility (PLATIM) and fly food production (Arthrotools).
Funding: This work was supported by French National Research Agency grants ANR-16-CE12-0006-01 (R.D.) and ANR-21-CE13-0037 (B.L.) and Damon-Runyon Cancer Research Foundation postdoctoral fellowship DRG 2438-21 (C.-H.C.). C.-H.C. is also supported by a National Institutes of Health grant R01-GM74108 to H. S. Malik.
Author contributions: Conceptualization: R.D., C.-H.C., and B.L.; Methodology: R.D., M.H., C.-H.C., B.H., and B.L.; Investigation: R.D., M.H., M.R., B.H., C.-H.C., and B.L.; Visualization: R.D., M.H., C.-H.C., B.H., and B.L.; Funding acquisition: R.D., C.-H.C., and B.L.; Supervision: R.D. and B.L.; Writing – original draft: B.L.; Writing – review and editing: R.D., M.H., B.H., C.-H.C., and B.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Fly stocks and plasmids are available upon request. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adh0037
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S6
References (*50*–*64*)
Movie S1

Submitted 3 February 2023; accepted 22 September 2023
10.1126/science.adh0037

WORKING LIFE

By **Eirini Tsekitsidou**

A second chance in science

It was a work day like any other—until a meeting with a suspiciously vague title popped up on my calendar for the next day. I immediately knew I was about to be laid off from my job as a consultant. When I had finished my Ph.D. 9 months earlier, I was convinced I was not cut out to be a scientist. So, I had pursued a career in the private sector instead. My self-confidence had just begun to recover—and now this. To make matters more complicated, I only had 3 months to find a new job or face deportation from the country I had called home for 11 years. Little did I know that this lowest point in my career would start me on my path to reclaiming my identity as a scientist.

Born and raised in Greece, I came to the United States for my undergraduate studies. I got used to life here and loved the country's diversity and wealth of opportunities. I did well academically, and after graduation I stayed to pursue a Ph.D. in cell and molecular biology. During graduate school, though, dead-end projects, rejected fellowship applications, and a lack of first-author publications convinced me I was not good enough for science. I had acquired some experience in management consulting and enjoyed learning about the business side of biotech, so I decided to pursue a career as a consultant instead—until the layoff.

"I only had 3 months to find a new job or face deportation."

Like any person who loses their job, I felt anger, hopelessness, and anxiety. These were made much worse by the visa clock ticking over my head. A 1-year extension of my student visa had enabled me to work, but I was only allowed 90 days of unemployment before I would have to leave the country. And not just any job would do: Per my visa, the position had to be directly related to my Ph.D. studies. Of course, I wanted to stay in my field as well—but I wasn't sure I could make it happen in time. I felt painfully unwelcome and rejected, alone, embarrassed, and vulnerable. I worried about how I would pay my rent, what would happen if I had to leave the country, when I would see my partner and my friends again.

After taking a few days to recover from the shock, I turned my attention to the immediate task: applying for any and every job I could find, hoping something would work out in time. Whereas in my previous post-Ph.D. job search I had ruled out science and academia—because of lack of confidence, not lack of desire—now I needed to explore every possible option, regardless of whether I thought the employer would take me.

That's when the surprises really started. It quickly became clear that private employers saw the need to sponsor a new work visa for me as a burden. Academia, on the other hand, was much more welcoming than I ever expected, based on what I saw as my lack of success as a grad student. When I contacted principal investigators whose labs I might be interested in working in, many responded enthusiastically. Even those in research areas unrelated to my Ph.D. were willing to consider inviting me into their group—not because I had the exact skills they were looking for, but because they saw my Ph.D. as evidence I was capable of learning and persevering. And I knew from experience that obtaining a visa while working at a university is relatively straightforward. Perhaps academic science had a place for me after all.

I have now begun a new chapter of my life as a postdoc, and I am beginning to regain my confidence. Some days are really hard, especially when an acquaintance sees me on campus—I'm at the same university where I completed my Ph.D.—and asks, "What are you doing here? I thought you graduated and worked at a fancy job now!" But most days are exciting, reminding me of the joys of reading papers, learning about new scientific findings, and, most important, working side by side with my new lab mates and enjoying their company. Academia certainly isn't perfect, and I feel for all the international researchers who may feel trapped in university labs because of their visa status. But I'm grateful I was forced to give science—and myself—a second chance. ■

Eirini Tsekitsidou is a postdoc at Stanford University. Send your career story to SciCareerEditor@aaas.org.